Studentisierte Permutationstests für verbundene und unverbundene 2-Stichprobenprobleme

Marius Placzek
Institut für Medizinische Statistik
Humboldtallee 32
Göttingen
marius.placzek@med.uni-goettingen.de

Frank Konietschke
Institut für Medizinische Statistik
Humboldtallee 32
Göttingen
fkoniet@gwdg.de

Markus Pauly
Mathematisches Institut
Universitätsstraße 1
Düsseldorf
markus.pauly@uni-duesseldorf.de

Zusammenfassung

Permutationstests sind als robuste nichtparametrische Testverfahren dafür bekannt, dass sie schon bei sehr kleinen Stichprobenumfängen den Fehler 1. Art akkurat kontrollieren. Eine wesentliche Annahme vieler Permutationsmethoden fordert dabei jedoch die Austauschbarkeit der Daten unter der Nullhypothese. Die Annahme der Austauschbarkeit ist jedoch sehr streng und in vielen Studien nicht erfüllt. Insbesondere impliziert sie Varianzhomogenität, das heißt gleiche Varianzen zwischen den Behandlungsgruppen. Hier werden wir sogenannte studentisierte Permutationstests sowohl für den unverbundenen als auch für den verbundenen 2-Stichproben-Fall vorstellen, die derartige Annahmen nicht benötigen. Da in SAS keine studentisierten Permutationstests für den 2-Stichprobenfall implementiert sind, werden wir zwei Makros zur Erweiterung der vorhandenen Routinen präsentieren.
Die Verfahren werden durch anschauliche Beispiele motiviert und illustriert, sowie die Verwendung der Makros demonstriert.

Schlüsselwörter: 2-Stichproben-Design, unverbundene Daten, verbundene Daten, studentisierte Permutationstests

1 Einleitung

In vielen Studien der Psychologie, Biologie oder Medizin werden zwei Stichproben erhoben, beispielsweise wenn zwei Gruppen von Patienten jeweils ein Medikament verabreicht wird (unverbundene Daten), oder aber, wenn Patienten zunächst das eine, dann das andere Medikament einnehmen und die Zielgröße jeweils gemessen wird (verbundene Daten).
Klassische parametrische Methoden zur Analyse solcher Daten, wie z.B. der verbundene und unverbundene t-Test, nehmen an, dass die Daten in den beiden Gruppen normalverteilt sind. Die Hypothese „kein Effekt“ wird dann über die Gleichheit der Er-

wartungswerte formuliert. Häufig ist diese Annahme jedoch nicht erfüllt, beispielsweise bei schiefen Verteilungen oder ordinalen Daten, sodass nichtparametrische Prozeduren anzuwenden sind. In der Praxis werden unverbundene Daten in der Regel mit dem Wilcoxon-Mann-Whitney Test (unter Annahme gleicher Varianzen) bzw. dem Brunner-Munzel-Test (ungleiche Varianzen) ausgewertet. Verbundene Messungen werden mit Hilfe des Wilcoxon-Matched-Pairs Tests (metrische Daten und gleiche Varianzen) bzw. des Munzel-Tests (metrische oder diskrete Daten und ungleiche Varianzen) ausgewertet. Hier werden Hypothesen über Gleichheit von Verteilungsfunktionen (bzw. keine Verschiebung im Shift-Modell) oder mit Hilfe eines so genannten relativen Effekts formuliert. Die zitierten Verfahren testen alle unterschiedliche Hypothesen und sind daher nicht ad-hoc vergleichbar, allen gemeinsam ist jedoch, dass sie bei Varianzheterogenität zu liberalen oder konservativen Testentscheidungen bei sehr kleinen Stichprobenumfängen tendieren.
Studentisierte Permutationstests sind im Gegenzug valide Inferenzverfahren auch unter der Annahme heterogener Varianzen zum Testen geeigneter Hypothesen bei kleinen Stichprobenumfängen. Des Weiteren können Konfidenzintervalle für die jeweiligen Behandlungseffekte mit den Verfahren berechnet werden, so wie dies z.B. für klinische Studien von internationalen Regulierungsbehörden explizit gefordert wird: „Estimates of treatment effects should be accompanied by confidence intervals, whenever possible…" (ICH, 1998, E9 Guideline, ch. 5.5, p.25) [1].
Im Folgenden werden zwei motivierende reale Datenbeispiele diskutiert. Anschließend wird das allgemeine Prinzip studentisierter Permutationstests für unverbundene und verbundene Zwei-Stichprobenprobleme vorgestellt. Die Güte der Prozeduren wird jeweils mit beispielhaften Simulationsergebnissen unterstrichen. Abschließend werden zwei SAS-Makros präsentiert, in denen die beschriebenen Methoden umgesetzt werden. Die Verwendung der SAS-Makros wird abschließend mit Hilfe der beiden Datenbeispiele genau beschrieben.

1.1 Motivierendes Beispiel 1 – unverbundene Daten

In einer Fertilitätsstudie mit 29 weiblichen Wistar - Ratten wurde die Wirkung einer Substanz (Verum) auf die Fertilität der Ratten untersucht. Dazu erhielten 12 Ratten ein Placebo, während den restlichen 17 Ratten das Medikament verabreicht wurde. Gemessen wurde dann jeweils die Anzahl der Implantationen (Einpflanzungen befruchteter Eier in die Gebärmutter) nach der Sektion der Tiere. Es handelt sich hier also um ein unverbundenes, unbalanciertes 2-Stichproben-Design mit Zähldaten. Es ist also zweifelhaft, ob die Annahme von normalverteilten, varianzhomogenen Daten zutreffend ist. Die Originaldaten sind aus [2] entnommen. Die Boxplots der beiden Gruppen (Placebo/Verum) in Abbildung 1 zeigen, dass die Daten schief verteilt sind. Es lässt sich schon eine leichte Tendenz erkennen, dass in der Verum-Gruppe höhere Anzahlen an Implantationen vorliegen. Das Beispiel wird später ausgewertet.

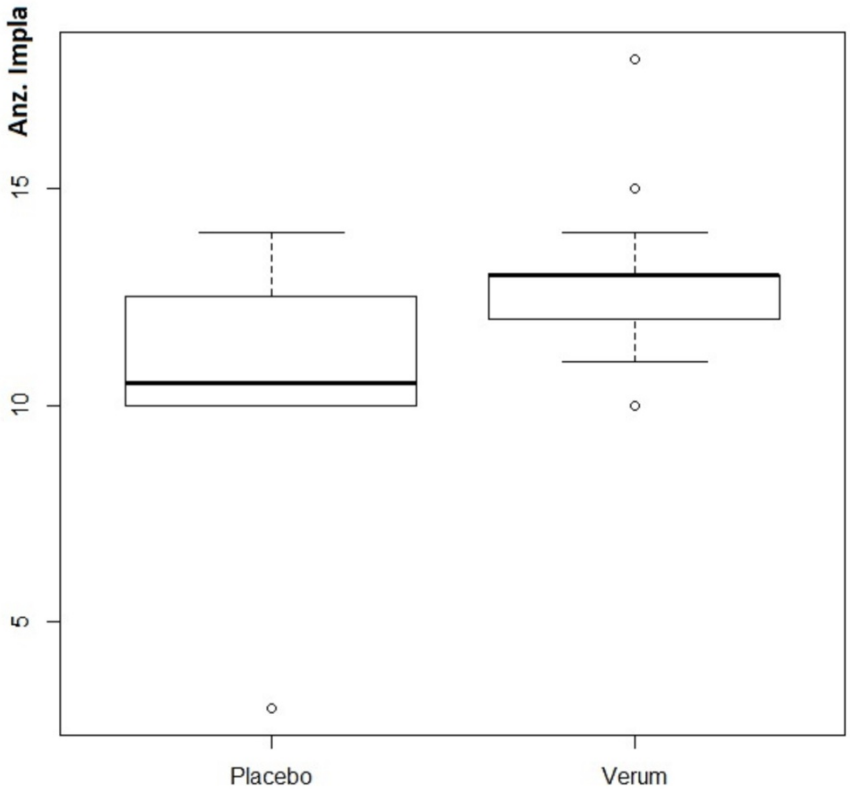


Abbildung 1: Boxplots zur Fertilitätsstudie

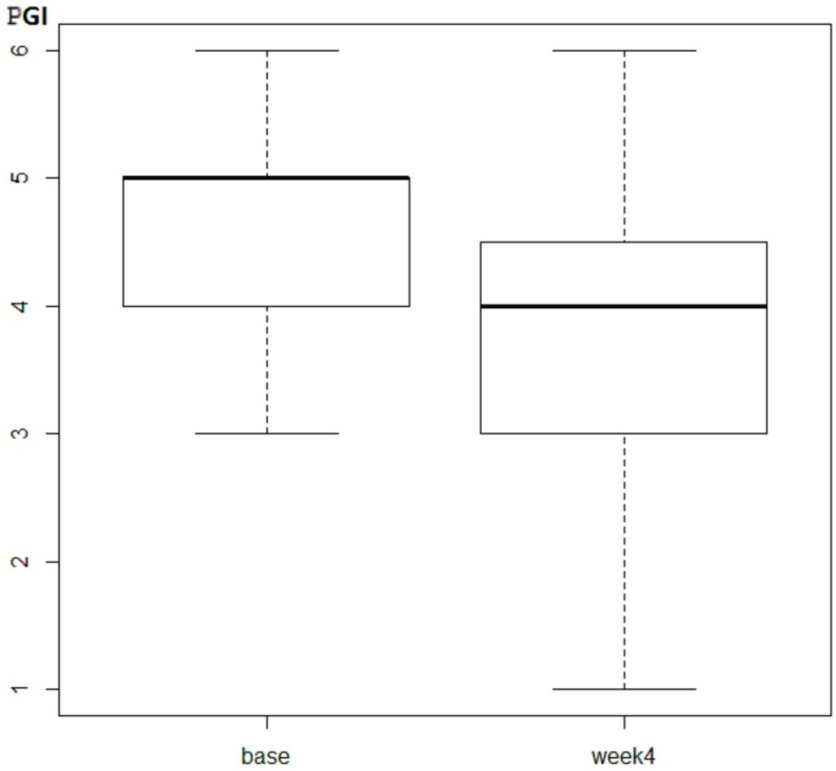


Abbildung 2: Boxplots zur Panikstörungsstudie

1.2 Motivierendes Beispiel 2 – verbundene Daten

Im Rahmen einer Angst-Studie (Panikstörung) mit n=15 Probanden wurde die Wirkung einer Bewegungstherapie auf das durch den Patienten bewertete klinische Befinden untersucht. Die Studienteilnehmer wurden zu Beginn der Therapie und 4 Wochen danach befundet, wobei als Responsegröße ein Score auf einer ordinalen Skala vergeben wurde. Je niedriger der Panikscore, desto besser ist das Befinden des Probanden. Jeder Patient wird somit wiederholt zu zwei Zeitpunkten gemessen. Die Originaldaten wurden [3] entnommen. Das Beispiel wird später sowohl mit einem mittelwertbasierten als auch mit einem rangbasierten Permutationstest ausgewertet.

2 Studentisierte Permutationstests

2.1 Zwei unverbundene Stichproben

Janssen (1997) [4] stellt einen studentisierten Permutationstest für das Behrens-Fisher-Problem im unverbundenen Fall vor. Für ausführlichere Theorie für den Zwei-Stichprobenfall siehe auch Janssen und Pauls (2003) [5], und für eine Erweiterung auf mehrfaktorielle Designs siehe Pauly et al. (2014) [14]. Gegeben seien unabhängig und identisch verteilte (u.i.v.) Zufallsvariablen $X_1, \dots, X_n$ sowie unabhängig von diesen u.i.v. $Y_1, \dots, Y_m$ mit $E(X_i) = \mu_1$ und $E(Y_i) = \mu_2$ sowie $Var(X_i) = \sigma_1^2$ und $Var(Y_i) = \sigma_2^2$. Dabei ist es zulässig, dass die Varianzen in den beiden Gruppen unterschiedlich sein können. Getestet werden soll nun die zweiseitige Hypothese $H_0: \mu_1 = \mu_2$. Die Welch's t-Test Statistik lautet

$$T = \frac{X_\cdot - Y_\cdot}{\sqrt{S_X^2/n + S_Y^2/m}}, \quad X_\cdot = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_\cdot)^2.$$

Die Verteilung dieser Teststatistik wird unter Nullhypothese mit einer t_ϑ-Verteilung mit ϑ Freiheitsgraden approximiert, wobei ϑ aus den Daten geschätzt wird. Die zweiseitige Hypothese H_0 wird verworfen, falls $|T| \geq t_{1-\alpha/2;\vartheta}$, wobei $t_{1-\alpha/2;\vartheta}$ das $(1-\alpha/2)$-Quantil der t_ϑ-Verteilung ist. Diese Methode ist jedoch anfällig auf die Verletzung der Normalverteilungsannahme und die Approximation durch die t-Verteilung wird erst für große Stichproben akkurat.

Der Permutationstest wird wie folgt durchgeführt:

1. Berechne die Teststatistik T
2. Führe Folgendes *nperm*-Mal durch (z.B. $nperm = 10000$):
- Permutiere den Vektor $(X_1, \dots, X_n, Y_1, \dots, Y_m)'$ zufällig und erhalte $X_1^*, \dots, X_n^*$ sowie $Y_1^*, \dots, Y_m^*$
- Berechne T_s^* aus den permutierten Daten
- Speichere T_s^* im Vektor $\boldsymbol{T}$
3. Ermittle aus dieser Permutationsverteilung das $(1-\alpha/2)$-Quantil $q_{1-\alpha/2;perm}$
4. Verwerfe H_0, falls $|T| \geq q_{1-\alpha/2;perm}$.

Das zugehörige $(1-\alpha)$-Konfidenzintervall für die Differenz der Erwartungswerte lässt sich dann wie üblich – nur hier mit dem Quantil der Permutationsverteilung – berechnen:

$$\left[(X_\cdot - Y_\cdot) - q_{1-\frac{\alpha}{2};perm}\sigma_{est}; (X_\cdot - Y_\cdot) + q_{1-\frac{\alpha}{2};perm}\sigma_{est}\right],$$

$$\sigma_{est}^2 = S_X^2/n + S_Y^2/m.$$

2.1.1 Simulationen

Janssen (1997) [4] präsentierte in seinem Paper eine Reihe von Monte-Carlo Simulationen. An dieser Stelle sei auszugsweise daraus eine Tabelle dargestellt, in der der klassische Welch-Test mit der Permutationsmethode verglichen wird (Tabelle 1). Dabei wird ein unbalanciertes Design mit sehr kleinen Stichprobenumfängen $(n = 4, m = 8)$ betrachtet. Das Setting ist so gewählt, dass man sich unter der Nullhypothese befindet, sodass die Kontrolle des Fehlers 1. Art (hier $\alpha = 0.0485$) verglichen werden kann. Die Daten werden von vier verschiedenen Verteilungen generiert und es gibt 5 Settings für die Varianzen. Es zeigt sich, dass der Welch-Test unter Normalverteilung noch gute Ergebnisse erzielt, auch bei leicht ungleichen Varianzen. Bei den schiefen Verteilungen werden die Testentscheidungen vor allem in Richtung *positive pairing* (große Varianz bei großer Stichprobe) recht konservativ, bei der Gleichverteilung in Richtung *negative pairing* (große Varianz bei kleiner Stichprobe) liberal. Der Permutationstest hält das Niveau im Großteil der Situationen recht gut ein, lediglich beim *negative pairing* wird er leicht liberal. Insgesamt (auch im Rest der Simulationsstudie) ist ein klarer Vorteil beim Permutationstest zu erkennen, gerade wenn nicht sichergestellt ist, ob die Daten normalverteilt und varianzhomogen sind.

Tabelle 1: Vergleich Welch-Test mit Permutationstest (10^6 Monte-Carlo Simulationen)

Type I error probability. $n_1 = 4$, $n_2 = 8$, $\mu_1 = \mu_2$, $\alpha = 0.0485$

Distribution	$\sigma_1^2 : \sigma_2^2$	1.2 : 1.0	1.1 : 1.0	1.0 : 1.0	1.0 : 1.1	1.0 : 1.2
Normal	$\phi_{n,\text{Welch}}$	0.0509	0.0495	0.0497	0.0488	0.0478
	$\phi_{n,\text{Perm}}$	0.0523	0.0497	0.0481	0.0456	0.0436
Log-normal	$\phi_{n,\text{Welch}}$	0.0453	0.0416	0.0379	0.0346	0.0316
	$\phi_{n,\text{Perm}}$	0.0555	0.0516	0.0486	0.0455	0.0429
Exponential	$\phi_{n,\text{Welch}}$	0.0518	0.0478	0.0442	0.0402	0.0377
	$\phi_{n,\text{Perm}}$	0.0556	0.0518	0.0481	0.0447	0.0426
Uniform	$\phi_{n,\text{Welch}}$	0.0663	0.0646	0.0627	0.0606	0.0589
	$\phi_{n,\text{Perm}}$	0.0538	0.0508	0.0482	0.0452	0.0430

Rangbasierte studentisierte Permutationstests für das nichtparametrische Behrens-Fisher Problem werden in [10], [11] und [12] entwickelt.

2.2 Zwei verbundene Stichproben

Konietschke und Pauly (2012) entwickeln in [6,7] studentisierte Permutationstests für zwei verbundene Stichproben. Sie beschreiben zwei Ansätze: Zum einen eine Methode, die auf den Mittelwerten basiert und die Daten komplett permutiert, ohne Rücksicht auf ihre Paarungen [6]. Zum anderen eine rangbasierte Methode, die den Ansatz von Janssen (1999) [13], bei der innerhalb der Paare permutiert wird [7], auf Bindungen erweitert.

2.2.1 Mittelwertbasierte Methode

Gegeben seien u.i.v. Zufallsvektoren $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$, $i = 1, \ldots, n$, mit $E(\boldsymbol{X}_1) = \boldsymbol{\mu} = (\mu_1, \mu_2)'$ und beliebiger Kovarianzmatrix $Cov(\boldsymbol{X}_1) = \boldsymbol{\Sigma}$. An der Hypothese ändert sich im Vergleich zum unverbunden Fall bei der mittelwertbasierten Methode nichts, d.h. wir testen weiter auf die Gleichheit der Erwartungswerte $H_0: \mu_1 = \mu_2$. Die klassische Teststatistik für den verbundenen t-Test lautet

$$T = \sqrt{n}\frac{(X_{1\cdot} - X_{2\cdot})}{\sqrt{\sigma_{est}^2}}, \qquad \sigma_{est}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{i1} - X_{i2} - (X_{1\cdot} - X_{2\cdot})\right)^2,$$

$$(\boldsymbol{X_{1\cdot}} - \boldsymbol{X_{2\cdot}}) = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}(X_{i1} - X_{i2}).$$

Bekanntermaßen ist T unter Nullhypothese exakt t_{n-1}-verteilt, falls die Differenzen normalverteilt sind, selbst bei beliebiger Kovarianzmatrix $\boldsymbol{\Sigma}$. Unter Nicht-Normalität approximiert man die Verteilung von T mit einer t_{n-1}-Verteilung. Die Nullhypothese wird verworfen, falls $|T| \geq t_{1-\alpha/2;n-1}$, wobei $t_{1-\alpha/2;n-1}$ das $(1-\alpha/2)$-Quantil der t_{n-1}-Verteilung ist. In zahlreichen Papern und Anwendungen wurde jedoch schon ge-

zeigt, dass unter Nicht-Normalität die Konvergenz von T zur asymptotischen Normalität sehr langsam abläuft und somit eine sehr große Stichprobe vonnöten ist – insbesondere bei schiefen Verteilungen [8].
Der Permutationstest, den Konietschke und Pauly (2012) [6] unter anderem vorschlagen, ist zunächst nicht intuitiv, da die paarige Struktur der Daten komplett im Permutationsalgorithmus ignoriert wird, weil alle $2n$ Beobachtungen permutiert werden.

Die Prozedur läuft wie folgt ab:

1. Berechne die Teststatistik T
2. Führe Folgendes nperm-Mal (z.B. $nperm = 10000$) durch:
   - Permutiere den Vektor $\boldsymbol{X} = (X_{11}, X_{12}, \dots, X_{n1}, X_{n2})'$ und erhalte $\boldsymbol{X}^* = (X^*_{11}, X^*_{12}, \dots, X^*_{n1}, X^*_{n2})'$
   - Berechne T_s^* aus den permutierten Daten $\boldsymbol{X}_i^* = (X_{i1}^*, X_{i2}^*)'$
   - Speichere T_s^* im Vektor $\boldsymbol{T}$
3. Ermittle aus dieser Permutationsverteilung das $(1-\alpha/2)$-Quantil $q_{1-\alpha/2;perm}$
4. Verwerfe H_0, falls $|T| \geq q_{1-\alpha/2;perm}$.

Das zugehörige $(1-\alpha)$-Konfidenzintervall für die Differenz der Mittelwerte lässt sich wieder wie üblich – nur hier mit dem Quantil der Permutationsverteilung – berechnen:

$$\left[(X_1 - X_2)_. - q_{1-\frac{\alpha}{2};perm}\sqrt{\sigma^2_{est}/n}; (X_{1.} - X_{2.}) + q_{1-\frac{\alpha}{2};perm}\sqrt{\sigma^2_{est}/n}\right].$$

2.2.2 Simulationen

Um den mittelwertbasierten Permutationstest mit dem verbundenen t-Test zu vergleichen, führen Konietschke und Pauly (2012) [6] eine Reihe von Simulationen in verschiedenen Settings durch. An dieser Stelle soll nun wieder ein kurzer Auszug daraus wiedergegeben werden, um zu zeigen, dass dieser kontraintuitive Permutationsansatz tatsächlich bei nichtaustauschbaren Daten valide Ergebnisse liefert.
Tabelle 2 zeigt ein solches Setting, in dem Daten mit den Marginalverteilungen

a) $F_1 = N(0,1)$ und $F_2 = N(0,2)$
b) $F_1 = N(0,1)$ und $F_2 = N(0,4)$
c) $F_1 = N(3,4)$ und $F_2 = \chi^2_3$
d) $F_1 = N(e^{0.5}, 3)$ und $F_2 = LN(0,1)$,

jeweils mit Korrelation $\rho \in (-1,1)$, simuliert wurden. Dabei wurde die Hypothese $H_0: \mu_1 = \mu_2$ getestet.
Die Ergebnisse zeigen, dass sowohl der verbundene t-Test als auch der studentisierte Permutationstest das Niveau recht akkurat einhalten, selbst bei nichtaustauschbaren Daten und sehr kleinen Stichproben ($n = 7$). Haben die marginalen Verteilungen jedoch extrem unterschiedliche Formen (d) und die Korrelation ist stark negativ, werden beide Tests leicht liberal.

Tabelle 2: Vergleich verbundener t-Test mit Permutationstest (10000 Simulationen)

Type-I error level ($\alpha = 5$ %) simulations for very small sample sizes ($n = 7$) with non-exchangeable distributions

Distribution	ρ	$T_{n,stud}$	Permutation test	Distribution	ρ	$T_{n,stud}$	Permutation test
(a)	−0.90	5.21	5.21	(c)	−0.90	6.14	6.63
	−0.50	4.84	4.96		−0.50	5.85	5.99
	−0.30	4.71	4.75		−0.30	5.44	5.80
	0.00	4.72	4.65		0.00	5.10	5.31
	0.30	5.15	5.19		0.30	5.36	5.67
	0.50	4.97	5.00		0.50	5.11	5.40
	0.90	4.90	4.88		0.90	5.50	5.53
(b)	−0.90	4.93	5.01	(d)	−0.90	6.94	7.51
	−0.50	4.82	4.71		−0.50	6.06	6.51
	−0.30	5.13	5.16		−0.30	5.53	5.94
	0.00	5.12	5.41		0.00	5.42	6.01
	0.30	5.18	5.25		0.30	5.04	5.71
	0.50	4.93	4.97		0.50	5.00	5.45
	0.90	5.09	5.35		0.90	5.91	6.13

2.2.3 Rangbasierte Methode

Gegeben seien u.i.v. Zufallsvektoren $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$, $i = 1, \dots, n$. An dieser Stelle nehmen wir jedoch lediglich an, dass das Modell marginal beschrieben werden kann, d.h. $X_{i1} \sim F_i$, i=1,2, mit den Randverteilungen F_1 und F_2.
Da dieses allgemeine Modell keinerlei Parameter beinhaltet, die zur Beschreibung eines Unterschieds zwischen den Verteilungen verwendet werden könnten, verwendet man die marginalen Verteilungen F_1 und F_2:

$$p = \int F_1 \, dF_2 = P(X_{11} < X_{22}) + 1/2P(X_{11} = X_{22}).$$

Der Effekt p wird auch relativer Marginaleffekt genannt [15]. Dieser ist sehr einfach zu interpretieren. Ist $p > 1/2$, so tendieren die Beobachtungen von F_2 zu größeren Werten als Beobachtungen der Verteilung F_1. Für $p = 1/2$ tendieren weder Beobachtungen von F_1 noch solche von F_2 zu größeren oder kleineren Werten. Folglich lässt sich die Hypothese „kein Behandlungseffekt“ über diesen Fall charakterisieren. Wir sind also daran interessiert die Nullhypothese $H_0: p = 1/2$ zu testen. Man beachte, dass aus $F_1 = F_2$ direkt $p = 1/2$ folgt, jedoch die Rückrichtung nicht gilt. Dies sieht man am Beispiel zweier Normalverteilungen mit gleichem Erwartungswert und unterschiedlichen Varianzen. Dort gilt zwar $p = 1/2$, jedoch nicht $F_1 = F_2$. Das Testproblem für $H_0: p = 1/2$ wird deshalb auch nichtparametrisches Behrens-Fisher-Problem genannt [9].
Munzel (1999) [8] schlug eine Testprozedur für $H_0: p = 1/2$ vor, die von vielen

Anwendern verwendet wird. Die zugehörige Teststatistik lautet

$$T = \sqrt{n}\frac{p_{est} - 1/2}{\sqrt{\sigma_{est}^2}},\ \ p_{est} = \frac{1}{2n}(R_{2.} - R_{1.}) + \frac{1}{2}\,, \qquad R_{1.} = \frac{1}{n}\sum_{i=1}^{n} R_{i1}\,,$$

$$Z_k = \frac{1}{n}\left(R_{i2} - R_{i2}^{(2)} - R_{i1} + R_{i1}^{(1)}\right),\ Z_{.} = \frac{1}{n}\sum_{k=1}^{n} Z_k\,,\ \sigma_{est}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_k - Z_{.})^2.$$

Dabei bezeichnet R_{i1} den Rang von X_{i1} unter allen $2n$ Beobachtungen und $R_{i1}^{(1)}$ den Rang von X_{i1} unter allen n Beobachtungen in Stichprobe 1.
Es lässt sich zeigen, dass die Teststatistik T unter Nullhypothese asymptotisch standardnormalverteilt ist, sodass die Nullhypothese verworfen wird, falls $|T| \geq z_{1-\alpha/2}$, wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung ist. Für kleine Stichproben schlägt Munzel (1999) eine Approximation mit einer t_{n-1}- Verteilung vor.
Simulationsstudien zeigen jedoch, dass diese Prozedur zwar das Niveau für Stichproben ≥ 20 sehr gut einhält, jedoch für kleinere Stichproben recht schnell sehr liberal oder konservativ wird, abhängig von negativer bzw. positiver Korrelation innerhalb der Paare X_{i1} und X_{i2}.
Konietschke und Pauly (2012) [7] geben eine Lösung für das nichtparametrische Behrens-Fisher-Problem bei verbundenen Daten an, indem sie nun innerhalb der Paare permutieren, immer wieder Munzels Teststatistik berechnen und so wiederum die kritischen Werte aus der Permutationsverteilung gewinnen. Die Prozedur wird wie folgt durchgeführt:

1. Berechne die Teststatistik T
2. Führe Folgendes nperm-Mal durch (z.B. $nperm = 10000$) (für kleine n lediglich alle Permutationen):
   - Permutiere die Paare $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$ und erhalte $\boldsymbol{X}_i^* = (X_{i1}^*, X_{i2}^*)'$
   - Berechne T_s^* aus den permutierten Daten
   - Speichere T_s^* im Vektor $\boldsymbol{T}$
3. Ermittle aus dieser Permutationsverteilung das $(1-\alpha/2)$-Quantil $q_{1-\alpha/2;perm}$
4. Verwerfe H_0, falls $|T| \geq q_{1-\alpha/2;perm}$.

Das zugehörige $(1-\alpha)$-Konfidenzintervall für den relativen Behandlungseffekt lässt sich - wieder mit dem Quantil der Permutationsverteilung – berechnen:

$$\left[p_{est} - q_{1-\frac{\alpha}{2};perm}\sqrt{\sigma_{est}^2/n}; p_{est} + q_{1-\frac{\alpha}{2};perm}\sqrt{\sigma_{est}^2/n}\right].$$

2.2.4 Simulationen

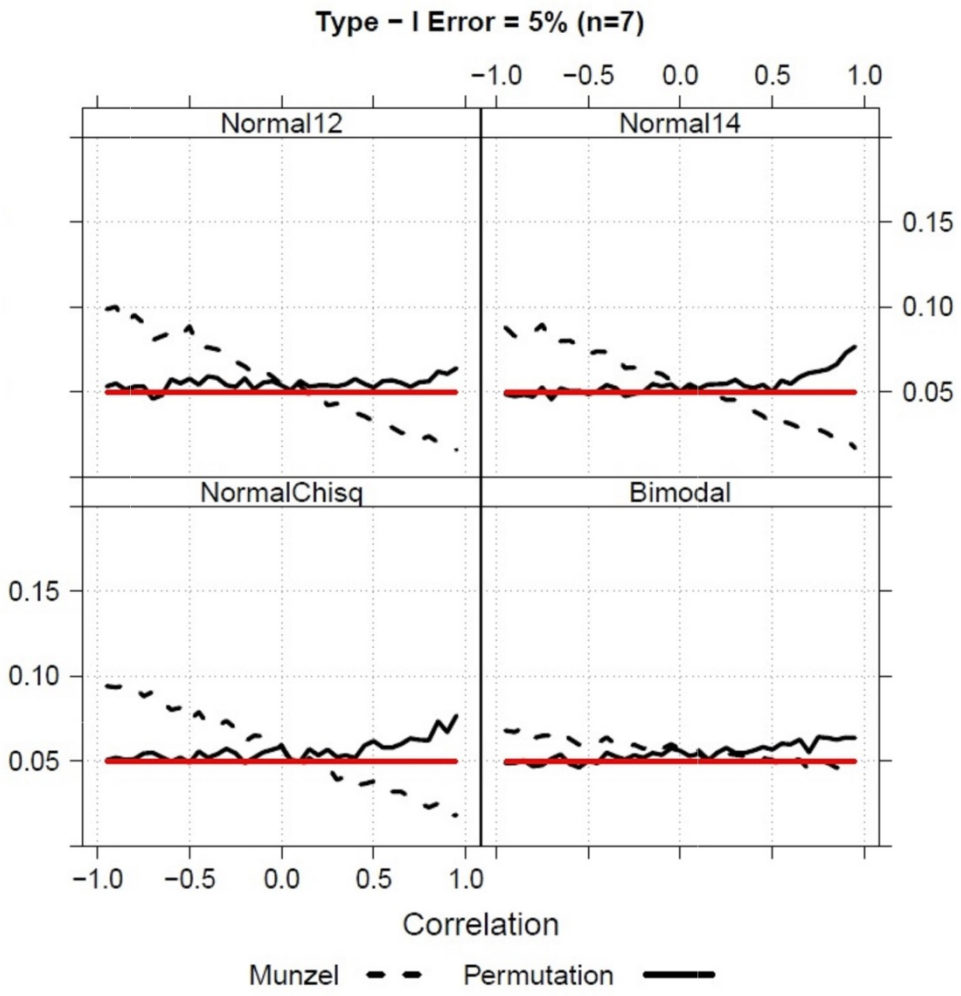


Abbildung 3: Simulierte Fehlerraten des Fehlers 1. Art bei nichtaustauschbaren Daten (Munzels Test vs. Permutationstest)

Die Kontrolle des Fehlers 1. Art und die Power des Verfahrens zur Aufdeckung fester Alternativen werden in Konietschke und Pauly [7] in extensiven Simulationsstudien untersucht. Abbildung 3 zeigt simulierte Raten für den Fehler 1. Art des Munzel Tests und des studentisierten Permutationstests für eine Stichprobengröße von $n = 7$. Die Anzahl der Simulationsdurchläufe beträgt $nsim = 10\,000$. Zu sehen sind vier Settings mit verschiedenen Randverteilungen F_1 und F_2, um ähnlich wie in Abschnitt 2.2.2 nichtaustauschbare Daten zu simulieren. Das Spektrum der Korrelationen ρ (x-Achse) liegt zwischen -1 und 1. Der Permutationstest zeigt eine viel bessere Kontrolle des nominalen 5%-Levels im Vergleich zu Munzels Test, der vor allem bei positiver oder negativer Korrelation konservativ bzw. liberal wird. Bei sehr stark positiver Korrelation wird der Permutationstest leicht liberal, was sich jedoch mit der relativ kleinen Stichprobengröße erklären lässt, wodurch lediglich $2^7 = 128$ Permutationen möglich sind.

3 SAS Makros

Da die in Abschnitt 2 zusammengestellten studentisierten Permutationsverfahren in dieser Form nicht in SAS verfügbar sind, haben wir zwei SAS Makros programmiert, mit Hilfe derer reale Daten leicht ausgewertet werden können. Die Verwendung und Syntax wird nun beschrieben. Zunächst wird das Makro `STUD_PERMU_TTEST` vorgestellt, das zur Analyse metrischer Daten (unverbunden oder verbunden) verwendet werden kann.

```
%STUD_PERMU_TTEST(DATA        = datensatz,
                  VAR         = abhängigevariable,
                  GROUP       = gruppierungsvariable,
                  SUBJECT     = subjektID,
                  ALTERNATIVE = alternative,
                  NPERM       = anzahlpermutationen,
                  PAIRED      = FALSE,
                  Alpha       = 0.05)
```

- DATA: Hier wird der Datensatz spezifiziert.
- VAR: Name der abhängigen Variable.
- GROUP: Name der gruppierenden Variable.
- SUBJECT: Name der Variablen, die die Subjekte identifiziert (nur bei verbundenen Daten notwendig).
- ALTERNATIVE: Gegen welche Alternative soll getestet werden? Zur Auswahl stehen “two.sided“, “greater“, “less“.
- NPERM: Anzahl der Permutationen (default=10000).
- PAIRED: Sind die Daten verbunden? TRUE/FALSE (default=FALSE).
- ALPHA: Numerischer Wert. Es werden $1-\alpha$ Konfidenzintervalle berechnet. (default alpha=0.05)

Die rangbasierten studentisierten Permutationstests für verbundene Daten sind im Makro `STUD_PERMU_MUNZEL` implementiert. Die Routine verwendet die gleiche Syntax wie das eben beschriebene Makro `STUD_PERMU_TTEST`. Es ist zur Analyse von verbundenen Daten geeignet und implementiert den studentisierten Permutationstest aufbauend auf Munzels Teststatistik. Daher fällt die Option `PAIRED` weg.

4 Auswertung der Beispiele

Die beiden Beispiele aus Abschnitt 1 werden nun mit Hilfe der soeben vorgestellten Makros analysiert.

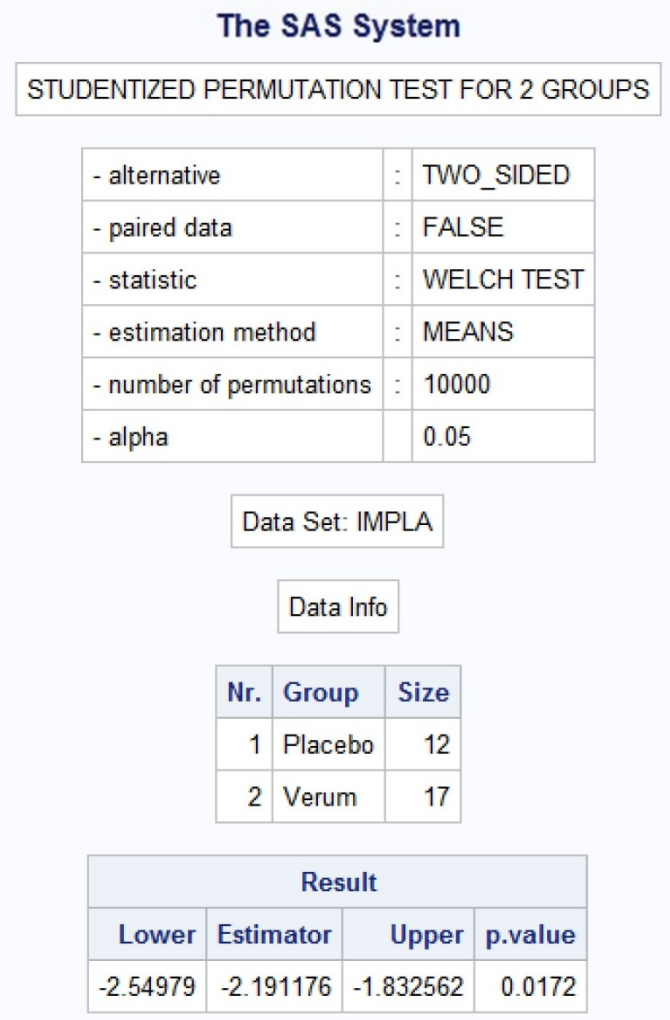
4.1 Auswertung Beispiel 1 – Fertilitätsstudie

Wie in Abschnitt 1.1 beschrieben handelt es sich hier um einen unverbundenen Datensatz. Daher eignet sich nur das erste Makro. Es wird wie folgt angewendet:

```
 DATA IMPLA;
input treatment$ impla;
datalines;
Placebo   3
…
Placebo  14
Verum    10
…
Verum    18
 ;
  RUN;

%STUD_PERMU_TTEST(DATA        = IMPLA,
                  VAR         = impla,
                  GROUP       = treatment,
                  ALTERNATIVE = "two.sided",
                  NPERM       = 10000,
                  ALPHA       = 0.05)       ;
```

Der Output aus dem *Results Viewer* ist in Abbildung 4 zu sehen. Nach einer kurzen Übersicht über die angewendete Prozedur werden noch einmal die Gruppen mit ihren Stichprobenumfängen aufgeführt, bevor in der abschließenden Tabelle der Schätzer für die Mittelwertdifferenz, sowie die obere und untere Schranke für das 95%-Konfidenzintervall angegeben werden, gefolgt vom p-Wert für die getestete Hypothese „kein Effekt“. Dieser liegt mit 0,0172 unter dem Signifikanzniveau von 0,05. Daher wird die Hypothese zu diesem Niveau abgelehnt. Die verwendete Substanz hat also tatsächlich einen Effekt auf die Fertilität der Wistar Ratten.



The SAS System

STUDENTIZED PERMUTATION TEST FOR 2 GROUPS

- alternative	:	TWO_SIDED
- paired data	:	FALSE
- statistic	:	WELCH TEST
- estimation method	:	MEANS
- number of permutations	:	10000
- alpha		0.05

Data Set: IMPLA

Data Info

Nr.	Group	Size
1	Placebo	12
2	Verum	17

Result			
Lower	Estimator	Upper	p.value
-2.54979	-2.191176	-1.832562	0.0172

Abbildung 4: Analyse der Fertilitätsstudie (SAS Output)

4.2 Auswertung Beispiel 2 – Panikstörungsstudie

Bei dem zweiten motivierenden Beispiel handelte es sich um einen verbundenen Datensatz. Daher sind beide Makros geeignet, um ihn zu analysieren. Zunächst werden die Daten entsprechend eingelesen. Dabei ist zu beachten, dass eine Variable die Subjekte identifiziert, sodass das Programm die zusammengehörigen Paare erkennen kann:

```
DATA PGI;
input pat time$ PGIscore;
datalines;
  1   base  6
  …
  15  base  6
  1   week4 4
  …
  15  week4 5
 ;
 RUN;
```

Im ersten Makro muss nun zusätzlich die Option `PAIRED=TRUE` ausgewählt werden:

```
%STUD_PERMU_TTEST(DATA        = PGI,
                  VAR         = PGIscore,
                  GROUP       = time,
                  SUBJECT     = pat,
                  ALTERNATIVE = "two.sided",
                  NPERM       = 10000,
                  PAIRED      = TRUE,
                  ALPHA       = 0.05)       ;
```

Abbildung 5 zeigt das Ergebnis.

The SAS System

STUDENTIZED PERMUTATION TEST FOR 2 GROUPS

- alternative	:	TWO_SIDED
- paired data	:	TRUE
- statistic	:	PAIRED T-TEST
- estimation method	:	MEANS
- number of permutations	:	10000
- alpha		0.05

Data Set: PGI

Data Info

Nr.	Group	Size
1	base	15
2	week4	15

Result			
Lower	Estimator	Upper	p.value
0	0.8666667	1.7333333	0.0276

Abbildung 5: Analyse der Panik-Studie (STUD_PERMU_TTEST)

Man kann jedoch auch das zweite Makro verwenden, um die Daten mit einem rangbasierten Permutationstest auszuwerten (Abbildung 6):

```
%STUD_PERMU_MUNZEL(DATA        = PGI,
                   VAR         = PGIscore,
                   GROUP       = time,
                   SUBJECT     = pat,
                   ALTERNATIVE = "two.sided",
                   NPERM       = 10000,
                   PAIRED      = TRUE,
                   ALPHA       = 0.05)        ;
```

The SAS System

STUDENTIZED PERMUTATION TEST FOR 2 GROUPS

- alternative	:	TWO_SIDED
- paired data	:	TRUE
- test statistic	:	MUNZEL
- estimation method	:	RANKS
- number of permutations	:	10000
- alpha	:	0.05

Data Set: PGI

Data Info

Nr.	Group	Size
1	base	15
2	week4	15

Result			
Lower	Estimator	Upper	p.value
0.1376735	0.2933333	0.4489932	0.0058

Abbildung 6: Analyse der Panik-Studie (STUD_PERMU_MUNZEL)

Beide Verfahren liefern einen p-Wert unter 5% (p.PERMU_TTEST=0,0276 und p.PERMU_MUNZEL=0,0058), d.h. die Nullhypothese „kein Effekt“ wird auch hier abgelehnt. Die Bewegungstherapie hat also einen positiven Effekt auf das Befinden.

5 Diskussion

Die in diesem Beitrag zusammengefassten studentisierten Permutationsverfahren bieten, wie die Simulationen zeigen, eine sehr gute Verbesserung der klassische Verfahren und können nun auch mit SAS auf reale Daten mit Hilfe der hier präsentierten Makros angewendet werden. Sie zeigen gerade bei kleinen Stichproben eine sehr gute Kontrolle des Fehlers 1. Art und eine zufriedenstellende Power zur Aufdeckung fester Alternati-

ven. Die Verfahren sind allesamt verteilungsfrei, d.h. es muss keine Annahme an die Verteilung der Daten gestellt werden, wie z.B. Normalität.

Literatur

[1] ICH, (1998): Statistical Principles for Clinical Trials. ICH, Guideline.

[2] E. Brunner, U. Munzel (2002): Nichtparametrische Datenanalyse. Springer-Verlag, New York, 2002.

[3] U. Munzel, E. Brunner (2002). An Exact Paired Rank Test. Biometrical Journal **44** 584–593.

[4] A. Janssen (1997): Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens Fisher problem. Statistics & Probability Letters **36**, 9-21.

[5] A. Janssen, T. Pauls (2003): How do bootstrap and permutation tests work? The *Annals of Statistics* **3**, 768–806.

[6] F. Konietschke, M. Pauly (2012). Bootstrapping and permuting paired t-test type statistics. *Statistics and Computing*. DOI 10.1007/s11222-012-9370-4.

[7] F. Konietschke, M. Pauly (2012): A studentized permutation test for the nonparametric Behrens-Fisher problem in paired data. Electronic Journal of Statistics **6**, 1358–1372.

[8] U. Munzel (1999). Nonparametric methods for paired samples. Statistica Neerlandica **53**, 277–286.

[9] E. Brunner, U. Munzel (2000). The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. Biometrical Journal **1,** 17–21.

[10] A. Janssen (1999). Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference*, **81** 71-93.

[11] A. Janssen (2001). Erratum: ``Testing nonparametric statistical functionals with applications to rank tests''. *Journal of Statistical Planning and Inference* **92**, 297.

[12] K. Neubert, E. Brunner (2007). A studentized permutation test for the nonparametric Behrens-Fisher problem. Computational Statistics and Data Analysis **51,** 5192 – 5204.

[13] A. Janssen (1999). Nonparametric symmetry tests for statistical functionals. *Mathematical Methods in Statistics* **8**, 320-343.

[14] M. Pauly, E. Brunner, F. Konietschke (2014). Asymptotic permutations for general factorial design. *Journal of the Royal Statistical Society – Series B*. In Press.

[15] E. Brunner, S. Domhof, & F. Langer (2002). Nonparametric Analysis of Longitudinal Data in Factorial ExperimentsJohn Wiley and Sons. New York.