Simulation der Power des t-Tests und des U-Tests bei ordinalen Daten

Anja Sander
Lorenz Uhlmann
Thomas Bruckner
Universität Heidelberg
Im Neuenheimer Feld 305
69120 Heidelberg
sander@imbi.uni-heidelberg.de

Zusammenfassung

Die in der Statistik wohl am häufigsten auftretende Testsituation ist der Vergleich von stetigen Merkmalen bei zwei unabhängigen Stichproben. Der dabei am meisten angewandte Test ist der t-Test. Er beruht auf der Annahme, dass das zu untersuchende Merkmal in der Grundgesamtheit normalverteilt ist. Im klinischen Alltag begegnet man oft ordinal skalierten Parametern (Scores), die zur Beurteilung der Befindlichkeit von Patienten erhoben werden. Für die Auswertung solcher ordinaler Daten bieten sich verschiedene statistische Testverfahren an:
Eine erste Möglichkeit bietet der Chi^2-Test, welcher prüft, ob zwei (oder mehr) Stichproben dieselbe Verteilung in einem nominal skalierten Parameter haben (Homogenitätstest). Er nützt also die ordinale Struktur der vorliegenden Daten nicht aus und ist daher im Falle ordinaler Daten eher konservativ. Eine Alternative, über welche die ordinale Struktur und zusätzlich auch evtl. vorhandene Bindungen berücksichtigt werden können, ist der Wilcoxon-U-Test. Dieser basiert auf Rängen und ist somit auch für stetige Daten geeignet. Er wird daher oft auch als nonparametrische Alternative zum t-Test verwendet. Ein Test der ebenfalls für die Auswertung ordinaler Daten geeignet ist, ist der Mantel-Haenszel-Test. Eine zunächst eher als ungeeignet erscheinende Art, diese Daten auszuwerten, ist die Anwendung des t-Tests, der, wie bereits oben angesprochen, Normalverteilung (und damit auch stetige Daten) voraussetzt. Es ist allerdings bekannt, dass der t-Test relativ robust ist gegen die Verletzung der Annahme von normalverteilten Daten: Es gibt auch Untersuchungen, die die Robustheit des t-Tests sogar bei ordinalen Merkmalen mit 3-5 Ausprägungen belegen.
In diesem Beitrag sollen diese konkurrierenden Tests hinsichtlich der Einhaltung des vorgegebenen Signifikanzniveaus, sowie der Power miteinander verglichen werden. Es werden dabei verschiedene Wege gezeigt, wie dieser Vergleich in SAS umgesetzt werden kann und welche Laufzeiten diese Wege jeweils mit sich bringen.

Schlüsselwörter: Ordinale Daten, Power, Simulation

1 Einleitung

Die in der Statistik wohl am häufigsten auftretende Testsituation ist der Vergleich von stetigen Merkmalen bei zwei unabhängigen Stichproben. Der dabei am meisten ange-

wandte Test ist der t-Test. Er beruht auf der Annahme, dass das zu untersuchende Merkmal in der Grundgesamtheit normalverteilt ist. Der t-Test wurde im Jahre 1908 von William Sealy Gosset eingeführt. Er arbeitete als Chemiker für die Guinness-Brauerei in Dublin (Irland) und entwickelte den t-Test als eine günstige und schnelle Möglichkeit, die Qualität des in der Brauerei hergestellten Bieres zu überwachen. Da Guinness seinen Mitarbeitern verbot, Ergebnisse zu publizieren, veröffentlichte Gosset seine Arbeit unter dem Pseudonym „*Student*“ [1,2].
Im klinischen Alltag begegnet man oft ordinal skalierten Parametern (Scores), die zur Beurteilung der Befindlichkeit von Patienten erhoben werden. Beispiele hierfür sind der NYHA-Score (4 Ausprägungen, wobei NYHA I in klinischen Studien realistischerweise nicht beobachtet wird) der New York Heart Association zur Beurteilung der Stärke der Herzinsuffizienz und der ASA-Score der American Society of Anesthesiologists, der den Allgemeinzustand eines Patienten beschreibt (6 Ausprägungen).
Für die Auswertung solcher ordinaler Daten bieten sich verschiedene statistische Testverfahren an:
Eine erste Möglichkeit bietet der Chi^2-Test, welcher prüft, ob zwei (oder mehr) Stichproben dieselbe Verteilung in einem nominal skalierten Parameter haben (Homogenitätstest). Er nützt also die ordinale Struktur der vorliegenden Daten nicht aus und ist daher im Falle ordinaler Daten eher konservativ.
Eine Alternative, über welche die ordinale Struktur und zusätzlich auch evtl. vorhandene Bindungen berücksichtigt werden können, ist der Wilcoxon-U-Test. Dieser basiert auf Rängen und ist somit auch für stetige Daten geeignet. Er wird daher oft auch als nonparametrische Alternative zum t-Test verwendet. Ein Test der ebenfalls für die Auswertung ordinaler Daten geeignet ist, ist der Mantel-Haenszel-Test.
Eine zunächst eher als ungeeignet erscheinende Art, diese Daten auszuwerten, ist die Anwendung des t-Tests, der, wie bereits oben angesprochen, Normalverteilung (und damit auch stetige Daten) voraussetzt. Es ist allerdings bekannt, dass der t-Test relativ robust ist gegen die Verletzung der Annahme von normalverteilten Daten [3]. Es gibt auch Untersuchungen, die die Robustheit des t-Tests sogar bei ordinalen Merkmalen mit 3-5 Ausprägungen belegen [4].

In diesem Beitrag sollen diese konkurrierenden Tests hinsichtlich der Einhaltung des vorgegebenen Signifikanzniveaus, sowie der Power miteinander verglichen werden. Es werden dabei verschiedene Wege gezeigt, wie dieser Vergleich in SAS umgesetzt werden kann und welche Laufzeiten diese Wege jeweils mit sich bringen.

2 Material und Methoden

Ziel unserer Arbeit war es, die erwähnten Tests (t-Test, Chi^2-Test, Mantel-Haenzel-Test und Wilcoxon Test) bzgl. der Einhaltung des Signifikanzniveaus und der Power zu vergleichen, wofür Monte-Carlo-Simulationen durchgeführt wurden. Dabei wurden jeweils zwei Gruppen (Kontroll- und Interventionsgruppe) bzgl. einer ordinalen

Zielgröße mit jeweils drei bzw. fünf Kategorien miteinander verglichen. Bzgl. die Kontrollgruppe werden zwei verschiedene Szenarien betrachtet:

- Symmetrisch verteilte Eventwahrscheinlichkeiten: Dies bedeutet dass die Wahrscheinlichkeit, dass ein Wert in eine der drei (bzw. fünf) Kategorien fällt, ausgeglichen ist, die Population also bzgl. der Ausprägungen eine homogene Verteilung aufweist.
- Schief verteilte Eventwahrscheinlichkeiten: Hierbei haben die Kategorien unterschiedliche Wahrscheinlichkeiten, wobei diesbezüglich die höchsten Werte an einem der beiden Ränder auftreten.

Diese beiden Szenarien werden betrachtet, um zu beurteilen, ob sich die verschiedenen Tests bzgl. Ihrer Güte (Einhaltung des Signifikanzniveaus, sowie die erreichte Power) hierbei unterschiedlich verhalten.
Für die Simulation des Fehlers erster Art werden die Wahrscheinlichkeiten für die Verteilung auf die Kategorien in der Interventionsgruppe gleich denen in der Kontrollgruppe gesetzt. Zur Analyse der Power werden für die Interventionsgruppe abweichende Eventwahrscheinlichkeiten angenommen. Genaue Angaben zur Wahl dieser Wahrscheinlichkeiten finden sich im Ergebnisteil.
Im Folgenden werden nun unterschiedliche Programmierschritte zur Implementierung in SAS aufgezeigt und hinsichtlich der Laufzeit verglichen.

2.1 Notation

Folgende Notation wird in den weiteren Abschnitten verwendet:

- 2 Gruppen: Kontrollgruppe (C) und Interventionsgruppe (I)
- n_C bzw. n_I: Fallzahl pro Gruppe
- k: Anzahl Kategorien mit Eventwahrscheinlichkeiten p_l (l = 1,…,k), sodass gilt $\sum_{l=1}^{k} p_l = 1$
- x_i^C bzw. x_j^I $(i = 1, \dots, n_C,\ j = 1, \dots, n_I)$: Ausprägungen der Zielvariable in der Kontroll- bzw. Interventionsgruppe
- m: Anzahl der Simulationsdurchläufe

2.2 Verwendete Tests:

Die hier verglichenen Tests gehören zu den Standardverfahren in der Statistik und sind in vielen Artikeln und einführenden Fachbüchern ausführlich beschrieben (s. z.B. Sachs [5]). Daher wird im Folgenden nur eine kurze Übersicht über die Verfahren gegeben. Dabei wird immer davon ausgegangen, dass die Nullhypothese besagt, die Verteilung der abhängigen Variable in den beiden Gruppen gleich ist. Die Alternativhypothese ist hierbei die Aussage, dass die Verteilungen in der Interventionsgruppe im Vergleich zur Kontrollgruppe unterschiedlich sind.

t-Test:

Der t-Test geht von Normalverteilung der Zielvariable in der Grundgesamtheit der zu untersuchenden Population aus. Entsprechend ist die Teststatistik (für den Vergleich zweier unabhängiger Stichproben) als

$$T = \frac{\bar{x}_C - \bar{x}_I}{S\sqrt{\frac{1}{n_C} + \frac{1}{n_I}}}$$

definiert. Mit $\bar{x}_C$ bzw. $\bar{x}_I$ wird das arithmetische Mittel über die Ausprägungen in den jeweiligen Gruppen und mit S die gepoolte Standardabweichung bezeichnet. Die Teststatistik ist approximativ t-verteilt mit $n_C + n_I - 2$ Freiheitsgraden. Die darin verwendeten Größen sind für ordinale Daten streng genommen eigentlich nicht sinnvoll anwendbar. Aufgrund der oben bereits angesprochenen Robustheit des Tests soll er als möglicher Konkurrent dennoch in den Vergleich aufgenommen werden.

Chi²-Test:

Der Chi²-Test ist für nominale Daten konzipiert und nutzt daher die ordinale Struktur nicht aus. Insofern werden für die Konstruktion der Teststatistik auch nur die beobachteten und die zu erwartenden Häufigkeiten in den einzelnen Kategorien verwendet. Die Teststatistik lässt sich über die Ausprägungen einer Kontigenztafel konstruieren. Die entsprechenden Einträge seien mit h_{gl} $(g = C, I,\ l = 1, \dots, k)$ bezeichnet, die unter der Nullhypothese zu erwartenden Einträge mit $\tilde{h}_{gl}$ $(g = C, I,\ l = 1, \dots, k)$. Die Teststatistik ergibt sich damit als

$$\chi^2 = \sum_{g \in \{C,I\}} \sum_{l=1}^{k} \frac{\left(h_{gl} - \tilde{h}_{gl}\right)^2}{\tilde{h}_{gl}}.$$

Es werden also die Differenzen zwischen den beobachteten und den erwarteten Häufigkeiten betrachtet. Die Richtung der Abweichungen geht aufgrund der Quadrierung verloren. Die Teststatistik ist approximativ Chi²-verteilt mit $k - 1$ Freiheitsgraden.

Mantel-Haenszel-Test:

Dieser Test verwendet in der Teststatistik die Pearson Korrelation (r), welche über die Ränge der Beobachtungen bebildet werden, zwischen der Gruppen- und der Zielvariable:

$$Q_{MH} = (n-1)r^2,$$

wobei mit n die Gesamtstichprobengröße bezeichnet wird. Die Teststatistik ist dann approximativ Chi²-verteilt mit einem Freiheitsgrad. Im SAS-Output wird dieser Test mit der Bezeichnung „Mantel-Haenszel Chi-Square“ gekennzeichnet (s. auch die SAS-Hilfeseite zur Prozedur FREQ). Das Vorgehen ist in den entsprechenden Artikeln genauer beschrieben [6, 7].

Wilcoxon-U-Test:

Der Wilcoxon-U-Test ist ein verteilungsfreier Test, nutzt aber das ordinale Niveau aus, indem er die Ränge aus den Stichproben betrachtet. Die Teststatistik kann wie folgt definiert werden:

$$T_W = \sum_{i=1}^{n_i} rg(x_{Ci})$$

bzw.

$$T_W = \sum_{i=1}^{n_i} rg(x_{Ii})$$

Die Ränge werden dabei über die gesamte Stichprobe gebildet. Für die Testentscheidung kann entweder eine exakte Verteilung oder approximativ eine Normalverteilung verwendet werden.

3 Ergebnisse

3.1 Simulation ordinaler Daten in SAS

DATA Step

Ein "einfacher" Weg besteht darin im DATA Step über die Funktion RAND gleichverteilte Daten zu generieren und diese entsprechend der vorgegebenen Wahrscheinlichkeiten in die Kategorien einzuordnen. Bei k=3 mit Wahrscheinlichkeiten p_1=0.1, p_2=0.3 und p_3=0.6 sieht der SAS Code wie folgt aus:

```
DATA b;
     DO i = 1 TO &n; * Fallzahl pro Gruppe
          x = RAND("uniform");
          *x = UNIFORM(0);
          cat = (x > 0.1) + (x > 0.4); *allgemeiner (x>p1)+(x>(p1+p2));
          OUTPUT;
     END;
RUN;
```

Die Funktion UNIFORM ist vergleichbar mit der Funktion RAND ("uniform"), jedoch hinsichtlich der Laufzeit langsamer.
Alternativ können die Zufallszahlen auf direktem Weg über die „Table"-Verteilung erzeugt werden: Über &pC wird der Vektor der Eventwahrscheinlichkeiten für die Kontrollgruppe, über &pI derjenige für die Interventionsgruppe übergeben.

```
%let m = 100000; /* Anzahl Simulationsdurchläufe */
%let n = 50; /* Fallzahl pro Gruppe */
```

```
%let pC = 0.1 0.1 0.8; /* Vektor für die Eventw'keiten in der
Kontrollgruppe */
%let pI = 0.2 0.3 0.5; /* Vektor für die Eventw'keiten in der
Interventionsgruppe */

DATA b (KEEP=m x g);
ARRAY probC [3] (&pC);
ARRAY probI [3] (&pI);
CALL streaminit (123); *seed;
DO m = 1 TO &m;  /* Anzahl Simulationsdurchläufe */
   DO i = 1 TO &n; /* Fallzahl pro Gruppe */
      x = RAND("Table", of probC[*]);
     g=1;
   OUTPUT;
   END;
   DO i = 1 TO &n;
     x = RAND("Table", of probI[*]);
     g=2;
   OUTPUT;
   END;
END;
RUN;
```

PROC IML

```
PROC IML;
CALL RANDSEED (1234);
x=RANDMULTINOMIAL (&m, &n, &pC);
x = j(&n,&m); y = j(&n,&m); /* allocate */
call randgen(x, "Table",&pC); call randgen(y, "Table",&pI); /* u ~
U[0,1] */
```

3.2 Auswertung

```
PROC TTEST DATA=b;
    BY samples;
    CLASS g;
    VAR x;
    ODS OUTPUT ttests=TTests(WHERE=(method='Pooled'));
RUN;

PROC FREQ DATA=b NOPRINT;
    ODS EXCLUDE fishersexact;
    BY samples;
    TABLE g*x / NOPRINT CHISQ (warn=none);*warn=none suppresses
warnings;
    OUTPUT OUT=chi CHISQ;
RUN;

PROC NPAR1WAY WILCOXON DATA=b NOPRINT;
    BY samples;
    CLASS g;
```

```
    VAR x ;
    OUTPUT OUT=pwil (keep=ngr samples p2_wil) ;
RUN;
```

Aus den jeweils berechneten p-Werten muss für die Bestimmung des Typ-I-Fehlers bzw. der Power die Testentscheidung abgeleitet und für die einzelnen Simulationsdurchläufe aufsummiert werden. Dies kann beispielsweise über einen DATA Step und anschließend PROC MEANS erfolgen.

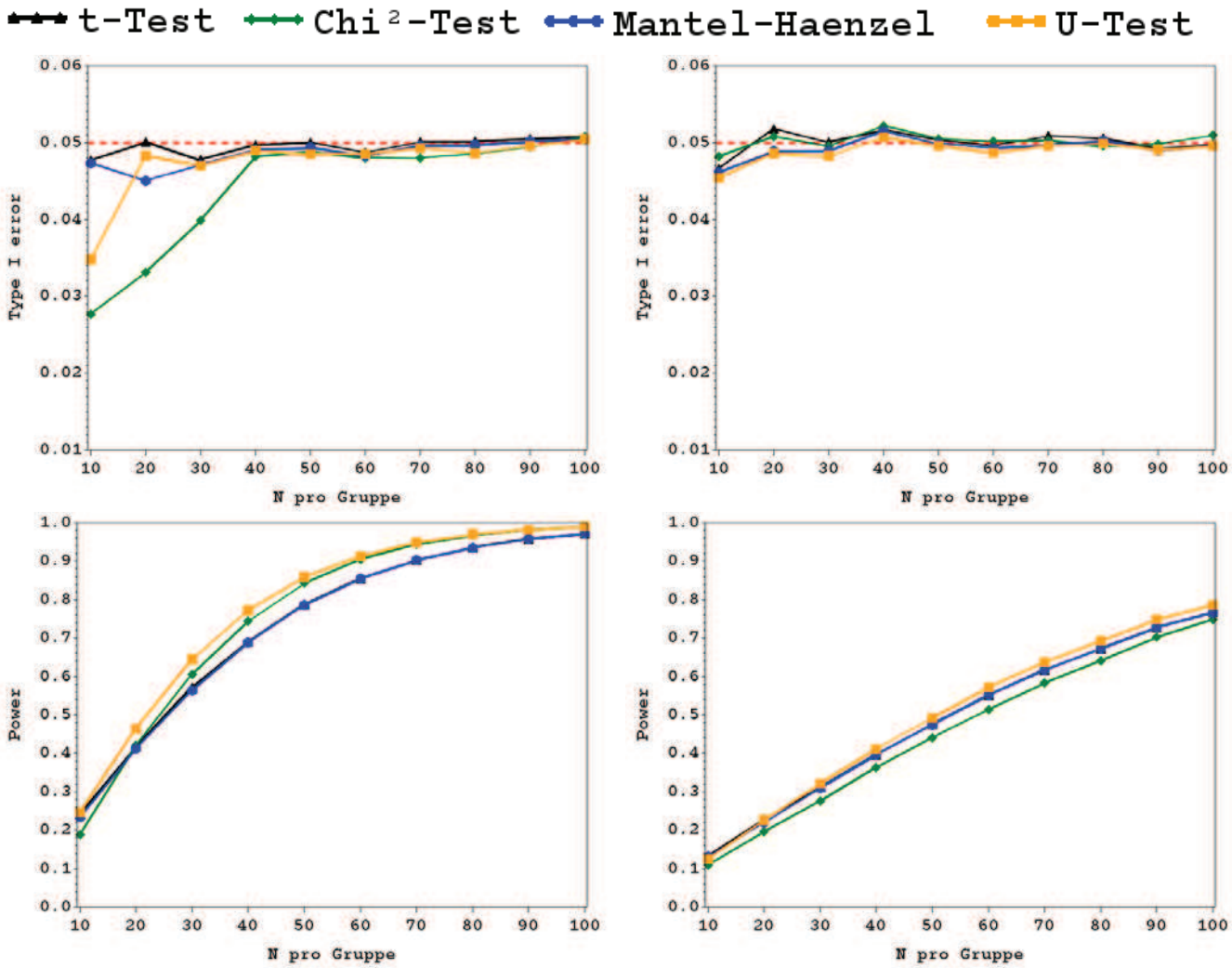
Alternativ in PROC IML

```
%let pC = {&pC};
%let pI = {&pI};

%let n =10;
%let m =100000;

PROC IML;
CALL RANDSEED (1234);

x = j(&n,&m); y = j(&n,&m); /* initiieren der Matrizen */
call randgen(x, "Table",&pC); call randgen(y, "Table",&pC); *
Simulation der kategoriellen Zufallszahlen;

/***************  t-TEST  ***************************/
meanX = mean(x); varX = var(x); /* Zeilenweise Verarbeitung */
meanY = mean(y); varY = var(y);
std=sqrt((varX+varY)/2);    /*    Berechnung    der    (gepoolten)
Standardabweichung */
B = loc(std=0); print B; std[B]=0.0001; /* Sonderfall std=0 -->
std=0.0001*/

t= (meanX - meanY) / (std*sqrt(1/&n + 1/&n)); /* Teststatistik */
alpha=0.05; RejectH0=(abs(t)>quantile("t",1-alpha/2, &n*2-2)); /*
Testentscheidung (0 oder 1) */

Prob = RejectH0 [:]; /* Ablehnungsrate */
print Prob;
QUIT;
```

3.3 Ergebnisse für 3 Kategorien

$$p_1^C = 0.1, p_2^C = 0.1, p_3^C = 0.8$$
$$p_1^I = 0.2, p_2^I = 0.3, p_3^I = 0.5$$

$$p_1^C = 0.3, p_2^C = 0.4, p_3^C = 0.3$$
$$p_1^I = 0.2, p_2^I = 0.3, p_3^I = 0.5$$



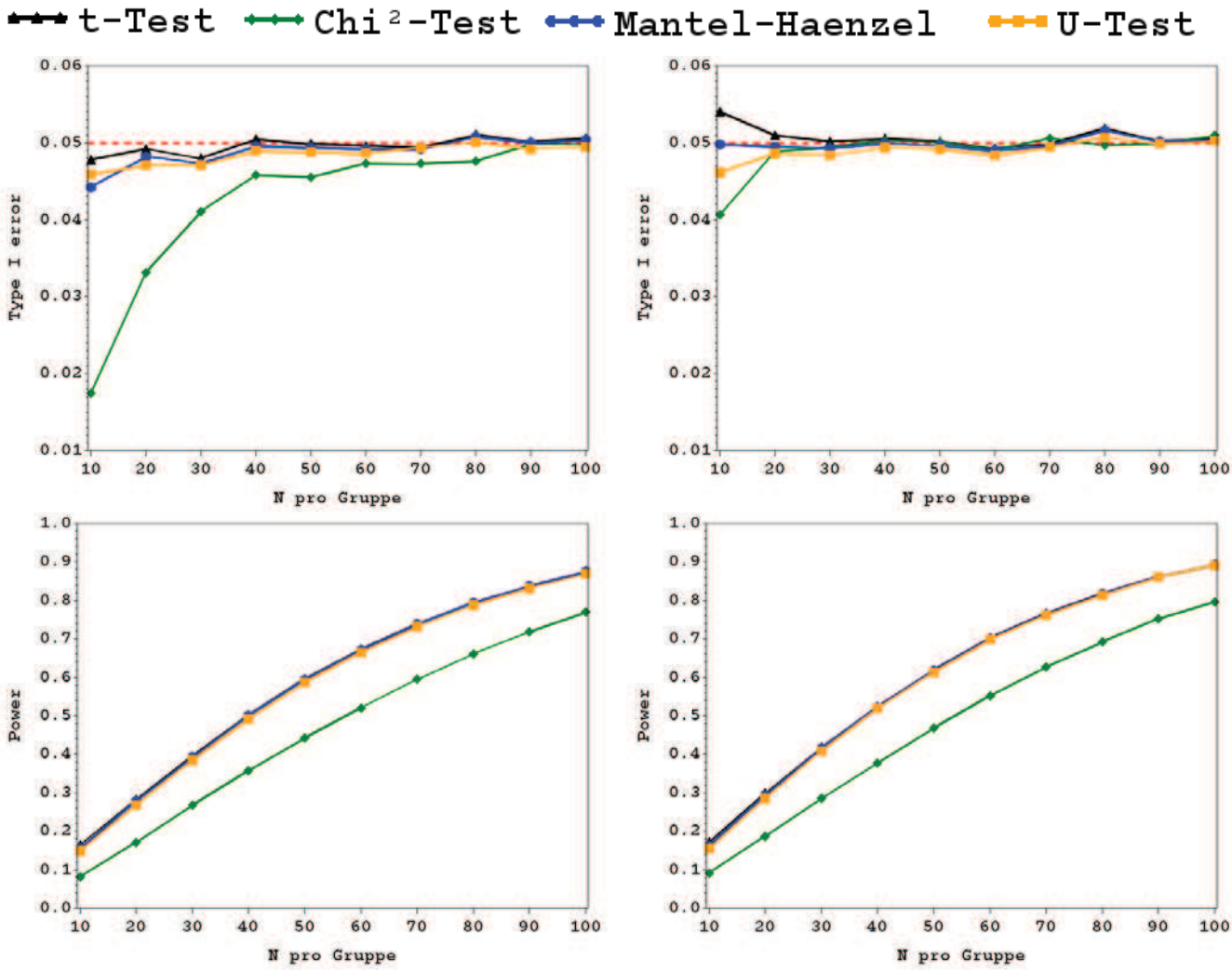
Generell ist festzuhalten, dass in den Szenarien mit drei Kategorien der Fehler erster Art von allen Tests bei allen Fallzahlen eingehalten wird (mit nur leichten Überschreitungen des nominellen Signifikanzniveaus). Der Chi²-Test, sowie der U-Test neigen bei kleinen Fallzahlen und schiefer Verteilung zu einem konservativen Verhalten (s. linke Grafiken). Auch der Mantel-Haenszel-Test ist davon (in etwas abgeschwächter Form) betroffen. Ab einer Fallzahl von 40 pro Gruppe liegen die Fehlerraten jedoch bei allen Tests gleichermaßen beim nominellen Signifikanzniveau. Bei symmetrischer Verteilung (s. rechte Grafiken) sind die Ergebnisse bzgl. des Typ-I-Fehlers für alle Tests und alle Fallzahlen sehr ähnlich. Lediglich bei einer Fallzahl von 10 pro Gruppe ist wieder ein leicht konservatives Verhalten von allen Tests zu beobachten, wobei hier nun der Chi²-Test am nächsten an der 5%-Linie liegt. Die Power der verschiedenen Tests liegt bei symmetrischer Verteilung relativ nahe beieinander, wobei der Chi²-Test die geringsten Werte mit sich bringt und der U-Test am besten abschneidet. Letzteres ist auch bei

schiefer Verteilung zu beobachten, während der Chi²-Test nun nur noch bei sehr geringer Fallzahl die geringsten Werte aufweist. In diesem Szenario sind bei höheren Fallzahlen die Ergebnisse von U- und Chi²-Test bzgl. der Power nahezu identisch, während t-Test und Mantel-Haenszel-Test über alle betrachteten Fallzahlen durchgehend auf demselben Niveau liegen.

3.4 Ergebnisse für 5 Kategorien

$p_1^C = 0.1, p_2^C = 0.1, p_3^C = 0.1,\ p_4^C = 0.1, p_5^C = 0.6$
$p_1^I = 0.2, p_2^I = 0.2, p_3^I = 0.1,\ p_4^I = 0.1, p_5^I = 0.4$

$p_1^C = 0.2, p_2^C = 0.2, p_3^C = 0.2,\ p_4^C = 0.2, p_5^C = 0.2$
$p_1^I = 0.1, p_2^I = 0.1, p_3^I = 0.2,\ p_4^I = 0.2, p_5^I = 0.4$



Beim Übergang auf Szenarien mit fünf Kategorien verändern sich besonders die Ergebnisse aus dem Chi²-Test. Dieser ist nun bei schiefer Verteilung (s. linke Grafiken) und bei geringer Fallzahl sehr konservativ. Erst ab 90 Beobachtungen pro Gruppe liegt das Niveau des Fehlers erster Art bei den vorgegebenen 5%. Alle anderen Tests weisen hingegen auch bei geringer Fallzahl Werte nahe dem nominellen Signifikanzniveau auf. Im Vergleich hierzu liegen bei symmetrischer Verteilung (s. rechte Grafiken) die Werte des Typ-I-Fehlers bzgl. aller Tests ab einer Fallzahl von 20 pro Gruppe sehr konstant beim 5%-Niveau. Bei 10 Beobachtungen pro Gruppe scheint der t-Test etwas antikonservativ

zu sein. Die Werte bzgl. der Power liegen für alle Tests sehr nahe beieinander, lediglich der Chi^2-Test führt hier zu niedrigeren Werten, was ihn insgesamt zum klaren Verlierer in den betrachteten Szenarien macht. Hier scheint sich das Ignorieren des ordinalen Skalenniveaus deutlich bemerkbar zu machen.

3.5 Laufzeit

Im Folgenden werden Laufzeiten für einzelne Programmschritte verglichen. Diese beziehen sich auf das Szenario mit 3 Kategorien und schiefer Wahrscheinlichkeitsverteilung $(p_1^C = 0.1, p_2^C = 0.1, p_3^C = 0.8; p_1^I = 0.2, p_2^I = 0.3, p_3^I = 0.5)$ mit $n_C = n_I = 50$ bei 100 000 Simulationsläufen. Dabei wird die „real time“ betrachtet, die aus fünf SAS-Aufrufen gemittelt wurde.
Die erste Möglichkeit, die Verwendung der Funktion UNIFORM und anschließende Kategorisierung im DATA Step, benötigt 23.04 Sekunden. Die alternative Möglichkeit über die Funktion RAND ("uniform") liegt fast identisch bei 23.06 Sekunden.
Die Generierung der Zufallszahlen über die RAND Funktion „Table“ benötigt 21.35 Sekunden und scheint damit leicht schneller zu sein. Bei dieser Größenordnung an Simulationsdurchläufen spielt die Generierung der Zufallszahlen hinsichtlich der Laufzeit keine entscheidende Rolle.

In Hinblick auf die Prozeduren ist es bei dieser großen Zahl an Simulationsdurchläufen erforderlich die automatisch erzeugten Ausgaben zu unterdrücken, da diese sonst das Output- bzw. Log-Fenster sprengen und die Rechenzeit wesentlich erhöhen.
Die Prozedur TTEST verfügt nicht über eine NOPRINT Option. Bei den anderen beiden Prozeduren FREQ und NPAR1WAY lassen sich die eigentlich automatisch produzierten Ausgaben über diese Option unterdrücken.

Bei 100 Simulationsdurchläufen (gewählt aufgrund von Machbarkeit) benötigt die Prozedur TTEST „normal“ (ohne Unterdrückung der Ausgaben) 53.65 Sekunden. Über folgende globale Optionen kann die Ausgabe unterdrückt werden:

```
ods graphics off ;
ods exclude all;
ods noresults;
```

Dadurch kann eine wesentliche Laufzeitverringerung erreicht werden, bezogen auf das obige Beispiel auf 0.11 Sekunden.
Zusätzlich können Notes unterdrückt werden über `OPTION NONOTES;` und die Anzahl der Fehlermeldungen im Log-Fenster über `OPTION ERRORS= ;` spezifiziert werden.

Beim Chi²-Test wird bei klein besetzten Zellen eine Warnmeldung ausgegeben, z.B.:

```
“WARNING: 33% of the cells have expected counts less than 5, for the
table of g by x. Chi-Square may not be a valid test.”
```

Diese kann über die Option (warn=none) im TABLE Statement unterdrückt werden.

Vergleich der Laufzeit zwischen PROC TTEST und IML und zwischen PROC TTEST mit und ohne WEIGHT Statement:

PROC IML benötigt für die Generierung der Zufallszahlen nur 0.25 Sekunden. Wenn man zusätzlich die Berechnung des t-Tests samt Ablehnungsrate bei $n_C = n_I = 50$ und m=100000 hinzunimmt, 8.69 Sekunden. Damit ist die Umsetzung über IML für den t-Test schneller als über DATA Step und PROC TTEST (Zum Vergleich: die Prozedur TTEST benötigt 24.65 Sekunden). Wobei hier angemerkt werden muss, dass größere Matrizen in PROC IML nicht ohne weiteres erzeugt werden können, da diese schnell die Kapazitäten des Arbeitsspeichers sprengen.

4 Diskussion

Die Betrachtung des Fehlers erster Art sowie der Power ist wichtiger Bestandteil der Evaluierung und des gegenseitigen Vergleichs von statistischen Tests. In diesem Beitrag haben wir verschiedene Wege einer entsprechenden Umsetzung in SAS dargestellt. Auch wenn es klarerweise noch weitere Möglichkeiten gibt, ordinale Daten im zwei-Gruppen-Vergleich zu analysieren, haben wir uns hier auf vier sehr häufig verwendete Verfahren beschränkt. Die Prozedur kann jedoch ohne größeren Aufwand auf weitere Tests erweitert werden.
In den Ergebnissen wurde besonders deutlich, dass der t-Test, obwohl er eigentlich gemäß seiner Annahmen nicht für ordinale Daten geeignet ist, sehr gut abschneidet, sprich den Fehler erster Art in nahezu allen Situationen einhält und bzgl. der Power zu konkurrierenden Tests vergleichbare Werte liefert. Der Chi^2-Test hingegen kann in den hier betrachteten Situationen nicht empfohlen werden, da er bei geringeren Fallzahlen zu eher konservativen Testentscheidungen neigt. Der Wilcoxon-U-Test kann als die beste Alternative gesehen werden, was aufgrund seiner Konstruktion für genau diese Datensituation nicht allzu sehr überraschend sein dürfte. Die Vorteile gegenüber bspw. des t-Tests sind teilweise aber erstaunlich gering. Insgesamt ist außerdem festzustellen, dass alle Tests in beinahe allen Situationen das vorgegebene Signifikanzniveau einhalten. Lediglich beim Szenario mit fünf Kategorien und symmetrischer Verteilung lag der t-Test leicht über der 5%-Marke. Die Laufzeit unterschied sich zwischen den verschiedenen Implementierungen nur sehr gering und war insgesamt auf einem sehr guten Niveau. Dennoch lässt sich erkennen, dass Proc IML den anderen Vorgehensweisen insgesamt überlegen ist (auch wenn diese Prozedur im Gegensatz zu allen anderen Prozeduren nicht zeilen- sondern matrizenorientiert ist und sich daher auch in der Programmierung unterscheidet), und sich daher besonders gut zur Umsetzung von Simulationsstudien eignet.

Literatur

[1] Seite „T-Test“. In: Wikipedia, Die freie Enzyklopädie. Bearbeitungsstand: 18. Juli 2014, 13:55 UTC. URL: http://de.wikipedia.org/w/index.php?title=T-Test&oldid=132259822

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

[3] K. Kubinger, D. Rasch und K. Moder (2009): Zur Legende der Vorraussetzungen des t-Tests für unabhängige Stichproben. *Psychologische Rundschau*, 60(1) , S. 26-27

[4] T. Heeren und R. d‘Agostino (1987): Robustness of the two independent samples t-Test when applied to ordinal scaled data. Statistics in Medicine, 6(1), S. 79-90

[5] L. Sachs (1991): Angewandte Statistik - Anwendung statistischer Methoden, Springer, Berlin, 7. Auflage.

[6] N. Mantel und W. Haenszel, (1959): Statistical Aspects of Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, 22(4), S. 719–748.

[7] R. J. Landis, E. R. Heyman und G. G. Koch (1978): Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests. *International Statistical Review*, 46(3), S. 237–254.