Maximal selektierte χ^2-Statistiken zur Untersuchung von Zero-Inflated Variablen

Bachelor-Thesis

im Studiengang Statistik

Eva-Maria Müntefering

München, Juli 2013

Betreuerin: Prof. Anne-Laure Boulesteix

Institut für Statistik
Ludwig-Maximilians-Universität München

Inhaltsverzeichnis

1 Einführung **5**

1.1 Einleitung 5

1.2 Übersicht 6

2 Bekannte statistische Tests **7**

2.1 Der χ^2-Unabhängigkeitstest 7

2.2 Der exakte Fisher-Test 8

2.3 Der Zweistichproben-t-Test 9

2.4 Der Wilcoxon-Rangsummen-Test 11

2.5 Der Kolmogorov-Smirnov-Test für Zweistichprobenprobleme . 12

2.6 Two-Part-Models 13

3 Maximal selektierte χ^2-Statistiken für ordinale Variablen **15**

4 Simulationsdesign **18**

4.1 Simulation 1 18

4.2 Simulation 2 19

4.3 Simulation 3, 4, 5 und 6 19

5 Simulationsergebnisse **20**

5.1 Simulation 1 21

5.2 Simulation 2 27

5.3 Simulation 3 30

5.4 Simulation 4 32

5.5 Simulation 5 und 6 36

6 Fazit **39**

7 Anwendung in der Praxis **42**

7.1 Überblick über die Daten 42

7.2 Ergebnisse 43

8 Zusammenfassung und Ausblick **47**

Abbildungsverzeichnis

1 Häufigkeit Maximaler Ablehnungsanteil Simulation 1 22
2 Häufigkeit Maximaler Ablehnungsanteil Simulation 2 27
3 Häufigkeit Maximaler Ablehnungsanteil Simulation 3 31
4 Häufigkeit Maximaler Ablehnungsanteil Simulation 4 34
5 Häufigkeit Maximaler Ablehnungsanteil Simulation 5 und 6 . . 39
6 Häufigkeiten Testpower > 0.8 40
7 Histogramm für kleine Daten 43
8 Anzahl Ablehnungen je Test für "Metastasiert" 44
9 Anzahl Ablehnungen je Test für "Tumorassoziiert". 46

Tabellenverzeichnis

1 Settings pro Durchgang für Simulation 1,3,4,5,6. 19
2 Settings pro Durchgang für Simulation 2. 20
3 Ablehnungsanteile Simulation 1 n=100 25
4 Ablehnungsanteile Simulation 1 n=40 25
5 Ablehnungsanteile Simulation 1 n=20 26
6 Ablehnungsanteile Simulation 2 n=40 30
7 Ablehnungsanteile Simulation 3 n=100 33
8 Ablehnungsanteile Simulation 3 n=50 33
9 Ablehnungsanteile Simulation 4 n=100 36
10 Ablehnungsanteile Simulation 4 n=50 37
11 Ablehnungsanteile Simulation 5 n=100 38
12 Ablehnungsanteile Simulation 6 n=100 38
13 Ablehnungen bei mind. einem Test, Metastasiert 45
14 Ablehnungen bei mind. einem Test, Tumorassoziiert 47

Zusammenfassung

In der Statistik gibt es eine Vielzahl an Tests, welche die Daten auf interessante Eigenschaften, wie zum Beispiel die Lage oder Mittelwerte, untersuchen sollen. Für viele dieser Tests werden bestimmte Annahmen vorausgesetzt. Es gibt Daten, zum Beispiel in der Medizin, in denen viele sehr kleine und einige große Werte vorherrschen. Es ist nicht voraussehbar, wie gut die Tests mit diesen Bedingungen umgehen. Will man die Daten auf einen Zusammenhang untersuchen, so verlangen viele Tests, zum Beispiel der Fisher-Test, die Angabe eines oder mehrerer Cutpoints, um die Variable in Kategorien zu unterteilen. Dieser ist vor Durchführung des Tests schwierig zu bestimmen. Ein Lösungsansatz dieses Problems ist eine Methode, die auf *maximal selektierten* χ^2*-Statistiken* basiert. In dieser Arbeit soll diese Methode mit den bereits bekannten Tests verglichen werden. Diese Tests sind der *Fisher-Test, t-Test, Wilcoxon-Test, Kolmogorov-Smirnov-Test* und der *Wilcoxon-* bzw. *t-Test* im sogenannten *Two-Part-Model*. Vergleichskriterium ist, wie gut die Tests einen Unterschied in Daten mit sehr vielen Nullen bzw. kleinen Daten erkennen. Dazu wurden mehrere Simulationen durchgeführt. Zuletzt wurde anhand eines Datensatzes über Krebspatienten untersucht, wie die Tests entscheiden. Ergebnis des Vergleichs ist, dass die Tests unterschiedlich gut auf verschiedene Situationen in den Daten reagieren. Die Methode mit den maximal selektierten χ^2-Statistiken scheint zusammen mit dem Wilcoxon-Test am besten für eine Detektierung eines Zusammenhangs zwischen einer binären und einer weiteren Variable, wie sie in den Simulationen gegeben sind, geeignet zu sein.

1 Einführung

1.1 Einleitung

Inhalt dieser Arbeit ist es, den Zusammenhang zwischen einer binären Größe und einer metrischen Größe mit vielen kleinen und einigen großen Werten bestmöglich zu untersuchen. Die der Statistik bereits bekannten Testmethoden bieten vielzählige Möglichkeiten vorliegende Daten auf interessierende Eigenschaften zu untersuchen. So gibt es zum Beispiel den *t-Test* oder *Gaußtest*, "um Hypothesen über den Parameter μ zu überprüfen" [Steland (2010), S.163], den *Wilcoxon-Test* (auch bekannt als Mann-Whitney-Test) oder mehrere χ^2*-Tests*. Zwei weitere Tests, die in dieser Arbeit benutzt werden, sind der *Exakte Fisher-Test* sowie der *Kolmogorov-Smirnov-Test*. Hinzu kommen,nach Lachenbruch (2001) und Lachenbruch (2002), sogenannte *Two-Part-Models*, sowie eine Methode, die auf *maximal selektierten* χ^2*-Statistiken* basiert (im Weiteren Verlauf teilweise als *Maxsel* bezeichnet) und von Boulesteix (2006) entwickelt wurde.

Die Durchführung statistischer Tests ist nicht immer problemlos möglich. Jedem Test liegen dabei unter Umständen bestimmte Annahmen zu Grunde, die vorausgesetzt werden. Ein Beispiel ist die Annahme, dass die Daten einer normalverteilten Grundgesamtheit entstammen. Diese greift insbesondere, beim t- bzw. Gaußtest. Es gibt Anwendungsfelder der Statistik, in denen gerade diese Annahmen nicht gemacht werden können. Ein Anwendungsfeld ist zum Beispiel die Medizin. Hier liegen häufig Daten vor, die viele kleine Werte und vor allem auch Nullen enthalten. Beispielsweise Daten, die den Effekt der Wirksamkeit eines Medikaments beschreiben (hat das Medikament keine Wirkung, so ist der entsprechende Wert 0) oder Daten, in denen es um interessierende (Blut-)Werte geht, welche bei gesunden Personen eher hoch sind, bei den kranken jedoch niedrig (oder umgekehrt). Größere Werte sind eher selten, sind deswegen jedoch nicht weniger wichtig. Die entsprechenden Daten folgen somit keinesfalls einer Normalverteilung. Sie sind eher stark linkssteil verteilt. Durch diese Verletzung der Annahmen ist nicht mehr sichergestellt, dass die angewandten Tests, hier speziell der t-Test, das

vorgegebene Signifikanzniveau α auch tatsächlich einhalten [Steland (2010), S.181].

Ein weiteres Problem ist die Herangehensweise an die Tests, insbesondere bei Tests, die normalerweise für kategoriale Daten gedacht sind, wie der *Fisher-Test*. Sind die Daten statt kategorial jedoch metrisch, so kann man die Daten vor der Berechnung bei einem Wert größer Null dichotomisieren, um so Kategorien zu erzeugen. Liegen Daten vor, bei denen alle Ausprägungen größer als Null sind, es jedoch viele verschiedene kleine Ausprägungen gibt, wird es schwierig, einen geeigneten Schwellenwert auszuwählen. Es gibt zwar die Möglichkeit, die Tests mit verschiedenen Schwellenwerten durchzuführen. Jedoch trifft man seine Entscheidung so aufgrund des kleinsten p-Wertes (*Fishing for Significance*), wovon dringend abzuraten ist. Für solche Daten sind Tests wie der *Fisher-Test* demnach eher ungeeignet.

Ziel der Arbeit ist es, die Maxsel-Methode mit den anderen, oben genannten Tests (t-Test, Wilcoxon-Test, Komogorov-Smirnov-Test, χ^2+t-Test, χ^2+Wilcoxon-Test) hinsichtlich ihrer Fähigkeit, Unterschiede in den Verteilungen zweier Gruppen zu erkennen, zu vergleichen. Dazu wurden zunächst verschiedene Daten simuliert, welche sich in den Anteilen der Nullen, den Parametern und somit auch in den Verteilungen unterscheiden. Danach wurde genauer betrachtet, wie sich die einzelnen Tests bezüglich der ausgegebenen p-Werte verhalten und dies kritisch bewertet. Weiter wurden die Tests auf einen realen Datensatz angewandt, welcher Daten über Krebspatienten enthält.

Die Simulation und die Berechnung der Ergebnisse erfolgte mit dem statistischen Programmpaket `R`, Version 3.0.1.

1.2 Übersicht

In Kapitel 2 werden die theoretischen Hintergründe der bereits bekannten Tests erläutert. Kapitel 3 beschäftigt sich mit der Methode der *Maximal selektierten χ^2-Statistik*. In Kapitel 4 wird zunächst das Simulationsdesign beschrieben, um danach in Kapitel 5 die Ergebnisse zu präsentieren. Inhalt von Kapitel 7 sind die Daten der Krebspatienten, die Durchführung der Tests

mit diesen und eine Aufstellung der gewonnenen Ergebnisse.

2 Bekannte statistische Tests

In den nachfolgenden Unterabschnitten werden die bereits bekannten Tests kurz vorgestellt. Dabei handelt es sich um den χ^2-*Unabhängigkeitstest*, den *exakten Fisher-Test*, den *t-Test*, den *Wilcoxon-Test* und den *Kolmogorov-Smirnov-Test*. Außerdem werden die *Two-Part-Models* vorgestellt.

2.1 Der χ^2-Unabhängigkeitstest

Der Test lässt sich auf der Grundlage kategorialer bzw. kategorisierter Daten X und Y berechnen. $(X_i, Y_i), i = 1, \ldots, n$ müssen dabei unabhängige Stichprobenvariablen sein. Diese sind in einer Kontingenztafel mit den Häufigkeiten h_{ij} und den Ausprägungen $(X = i, Y = j)$ darstellbar. Getestet wird, ob X und Y unabhängig sind. Die Nullhypothese lässt sich deshalb schreiben als

$$H_0 : P(X = i, Y = j) = P(X = i)P(Y = j)$$

oder vereinfacht ausgedrückt

$$\begin{aligned} H_0 : \pi_{ij} &= \pi_{i.}\pi_{.j} \qquad \forall\, i, j, \\ P(X = i, Y = j) &= \pi_{ij}, \\ P(X = i) &= \pi_{i.} \\ \text{und } P(Y = j) &= \pi_{.j}. \end{aligned}$$

Die Gegenhypothese lautet entsprechend

$$H_1 : P(X = i, Y = j) \neq P(X = i)P(Y = j),$$

für mindestens ein Paar (i, j). Wichtig zur Berechnung ist, dass die Randhäufigkeiten $h_{i.}$ und $h_{.j}$ gegeben sind. Die entsprechenden Randwahrscheinlichkeiten lassen sich durch $\hat{\pi}_{i.} = \frac{h_{i.}}{n}$ und $\hat{\pi}_{.j} = \frac{h_{.j}}{n}$ schätzen, sowie unter der Nullhypothese der Unabhängigkeit $\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j}$. Die letztendlich zu berech-

nende Teststatistik lautet

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}, \qquad \tilde{h}_{ij} = \frac{h_{i.}h_{.j}}{n}$$

mit

$$\chi^2 \overset{H_0}{\sim} \chi^2((k-1)(m-1)).$$

Der Ablehnungsbereich der Nullhypothese beim χ^2-Unabhängigkeitstest ist

$$\chi^2 > \chi^2_{1-\alpha}((k-1)(m-1)),$$

wobei entsprechende Quantile einer Tabelle zu entnehmen sind [Fahrmeir *et al.* (2010), S.467f.].

Im Fall, dass die Anzahl der Freiheitsgrade > 1 beträgt, können keine gerichteten Hypothesen formuliert werden. Die Voraussetzungen für die Durchführung des Tests sind

- Weniger als $\frac{1}{5}$ aller Zellen der Kreuztabelle haben eine erwartete Häufigkeit < 5.
- Keine Zelle weist eine erwartete Häufigkeit < 1 auf [Leonhart (2009), S.207,210].

2.2 Der exakte Fisher-Test

Sind die Voraussetzungen des χ^2-Tests nicht erfüllt oder sind die Stichprobenumfänge n_1 und n_2 nicht groß genug, um approximierte Verfahren anwenden zu können, so sollte stattdessen der exakte Fisher-Test durchgeführt werden. Die Stichproben $\mathbf{X} = (X_1, \ldots, X_{n_1})$ und $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})$ sind unabhängig. Auch dieser Test ist für kategoriale Daten vorgesehen. Interessierende Größen sind die Wahrscheinlichkeiten

$$p_1 = P(X_i = 1),$$
$$p_2 = P(Y_i = 1)$$

und die Nullhypothese

$$H_0 : p_1 = p_2$$

vs.

$$H_1 : p_1 \neq p_2.$$

Um eine Testgröße zu konstruieren, wird auf die beiden Zufallsvariablen $X = \sum_{i=1}^{n_1} X_i$ und $Y = \sum_{i=1}^{n_2} Y_i$, sowie die bedingte Verteilung von X gegeben $X + Y$, unter H_0, gegeben als

$$\begin{aligned} P(X = t_1 | X + Y = t_1 + t_2 = t) &= \frac{P(X = t_1)P(Y = t - t_1)}{P(X + Y = t)} \\ &= \frac{\begin{pmatrix} n_1 \\ t_1 \end{pmatrix} \begin{pmatrix} n_2 \\ t - t_1 \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ t \end{pmatrix}}, \end{aligned}$$

zugegriffen. Diese entspricht unter H_0 der hypergeometrischen Verteilung $H(n_1 + n_2, n_1, t)$. Um entscheiden zu können, ob H_0 abgelehnt werden kann oder nicht, wird der kritische Bereich $K = \{0, \ldots, k_u - 1\} \cup \{k_o + 1, \ldots, t\}$ aus

$$P(X > k_o | X + Y = t) \leq \alpha/2$$

und

$$P(X < k_u | X + Y = t) \leq \alpha/2$$

so bestimmt, dass k_u und k_o die größte bzw. kleinste Zahl ist, die die jeweilige Niveaubedingung einhält. H_0 wird abgelehnt, falls $X = t_1 \in K$ gilt [Toutenburg und Heumann (2008), S.153f.].

2.3 Der Zweistichproben-t-Test

Der *t-Test* beschäftigt sich mit Hypothesen über den Parameter μ zweier normalverteilten Variablen $X \sim N(\mu_X, \sigma_X^2)$ und $Y \sim N(\mu_Y, \sigma_Y^2)$. Es wird vorausgesetzt, dass die Stichproben $(X_1, \ldots, X_{n_1})$ und $(Y_1, \ldots, Y_{n_2})$ unabhängig

sind. Das zu testende Hypothesenpaar ist

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

Man unterscheidet bei der Testberechnung drei verschiedene Fälle [Toutenburg und Heumann (2008), S.142f.]:

- **Die Varianzen sind bekannt**: Falls die Varianzen bekannt sind lautet die Prüfgröße

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{\sqrt{n_1\sigma_X^2 + n_2\sigma_Y^2}} \cdot \sqrt{n_1 \cdot n_2} \overset{H_0}{\sim} N(0, 1).$$

  Unter der Nullhypothese ist diese standardnormalverteilt. n_i ist der Umfang der i-ten Stichprobe mit $i = 1, 2$. σ^2 entspricht der Varianz der jeweiligen Stichprobe (X oder Y). $\bar{X}$ steht für die Schätzung des unbekannten Erwartungswertes anhand des arithmetischen Mittels der Stichprobenwerte und es gilt $\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$. Die Schätzung von $\bar{Y}$ folgt analog. H_0 wird abgelehnt, falls $|T| > z_{1-\alpha/2}$ gilt [Toutenburg und Heumann (2008), S.132f., S.143].

- **Die Varianzen sind unbekannt, aber gleich**: In diesem Fall lautet die Prüfgröße

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \overset{H_0}{\sim} t(n_1 + n_2 - 2)$$

  mit

$$S = \sqrt{\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}}.$$

  S^2 ist die gemeinsame Varianz der Stichproben X und Y, welche durch die gepoolte Stichprobenvarianz geschätzt wird. S_X^2 entspricht der geschätzten Varianz von X. S_Y^2 entsprechend der von Y. $\bar{X}$ und $\bar{Y}$ werden berechnet wie im Fall der bekannten Varianzen. Die Prüfgröße besitzt unter der Nullhypothese eine Student'sche t-Verteilung mit $n_1 + n_2 - 2$

Freiheitsgraden. Falls $|T| > t_{n-1;1-\alpha/2}$ gilt, wird H_0 abgelehnt [Toutenburg und Heumann (2008), S.135, S.143].

- **Die Varianzen sind unbekannt und ungleich**: Falls $\sigma_X^2 \neq \sigma_Y^2$ gilt, gibt es keine exakt bestimmbare Testgröße, sondern nur die Näherungslösung

$$T(\mathbf{X}, \mathbf{Y}) = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \overset{H_0}{\sim} t(\nu)$$

mit

$$\nu = \left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2 / \left(\frac{(s_x^2/n_1)^2}{n_1 - 1} + \frac{(s_y^2/n_2)^2}{n_2 - 1}\right)$$

ganzzahlig gerundet. Die Nullhypothese wird abgelehnt, falls $|T| > t_{n-1;1-\alpha/2}$ gilt [Toutenburg und Heumann (2008), S.135, S.145].

2.4 Der Wilcoxon-Rangsummen-Test

Dieser nonparametrische Test wird angewandt, falls die Daten nicht-normalverteilt sind und ist somit eine Alternative zum *t-Test*. Die Idee des Tests ist, dass die Werte beider zu vergleichenden Stichproben gut durchmischt sein sollten, falls die Nullhypothese

$$H_0 : x_{med} = y_{med}$$

gilt. x_{med} steht für den Median der Stichprobe X, y_{med} analog für den Median der Stichprobe Y. Voraussetzung des Tests ist, dass die Verteilungsfunktionen beider Stichproben dieselbe Form besitzen. Um die Teststatistik aufzustellen, müssen zunächst die Ränge aller Beobachtungen der gepoolten Stichprobe, d.h. X und Y zusammen, bestimmt werden. Bei Bindungen, d.h. wenn ein Rang doppelt vorkommt, wird der Durchschnittsrang berechnet. Die Hypothesenpaare sind

a $H_0 : x_{med} = y_{med}$ vs. $H_1 : x_{med} \neq y_{med}$

b $H_0 : x_{med} \geq y_{med}$ vs. $H_1 : x_{med} < y_{med}$

c $H_0 : x_{med} \leq y_{med}$ vs. $H_1 : x_{med} > y_{med}$.

In dieser Arbeit ist das Hypothesenpaar (a) wichtig. Die letztendliche Teststatistik lautet

$$T_W = \sum_{i=1}^{n_x} rg(X_i).$$

n_x ist die Anzahl der Werte in Stichprobe X. Es werden die Ränge der Beobachtungen, die ursprünglich aus X stammen, aufsummiert. Die zugehörigen Ablehnungsbereiche lauten:

a $T_W > w_{1-\alpha/2}(n, m)$

b $T_W < w_{\alpha}(n, m)$

c $T_W > w_{1-\alpha}(n, m)$

mit $w_{\tilde{\alpha}}$ ist $\tilde{\alpha}$-Quantil der tabellierten Verteilung [Fahrmeir *et al.* (2010), S.459f.].

2.5 Der Kolmogorov-Smirnov-Test für Zweistichprobenprobleme

Für den Kolmogorov-Smirnov-Test sind die zwei unabhängigen Stichproben $X_1, \ldots, X_{n_1}$ und $Y_1, \ldots, Y_{n_2}$ gegeben. Die zugehörigen Zufallsvariablen sind $\mathbf{X} \sim F$ und $\mathbf{Y} \sim G$. F ist die empirische Verteilungsfunktion von $\mathbf{X}$, G die von $\mathbf{Y}$. Es soll geprüft werden, ob sich die Verteilungen beider Zufallsvariablen signifikant unterscheiden. Dazu wird die Nullhypothese

$$H_0 : F(t) = G(t)$$

gegen die Alternativhypothese

$$H_1 : F(t) \neq G(t)$$

getestet. Um eine Entscheidung zu treffen, werden die Differenzen zwischen beiden empirischen Verteilungsfunktionen gebildet. Die zugehörige Teststa-

tistik ergibt sich somit als

$$K = \max_{t \in R} |\hat{F}(t) - \hat{G}(t)|.$$

Für die Praxis genügt es jedoch, nur den Abstand für $t \in S$ zu bestimmen. S ist die (gepoolte) Stichprobe $S = \mathbf{X} \cup \mathbf{Y}$ und es gilt die Teststatistik

$$K = \max_{t \in S} |\hat{F}(t) - \hat{G}(t)|.$$

Der Ablehnbereich für die Nullhypothese ist

$$K > k_{n_1,n_2;1-\alpha},$$

wobei $k_{n_1,n_2;1-\alpha}$ aus entsprechenden Tabellen entnommen werden kann [Toutenburg und Heumann (2008), S.172].

2.6 Two-Part-Models

Wie in der Einführung bereits kurz erwähnt, gibt es "Two-Part-Models", welche vor allem benutzt werden, um die Klumpung der Daten bei der Null zu berücksichtigen. Folgende Erklärung dieser Modelle wurde sinngemäß aus [Lachenbruch (2002)] und [Lachenbruch (2001)] übernommen.

Wie der Name sagt, bestehen sie aus zwei Modellen. Die Responsevariable hat bei dieser Art von Modell die Form $y = (x, d)$ mit $d = 1$, falls y beobachtet wurde oder positiv ist, und $d = 0$, falls y entsprechend fehlt oder $y = 0$. Daraus folgt, dass der Response den Wert x annimmt, also $y = x$, falls $d = 1$ und ansonsten nicht definiert ist. Die Wahrscheinlichkeitsfunktion dieser "Two-Part-Models" der i-ten Gruppe lautet

$$f_i(x, d) = \left[p_i^{1-d} \left\{ (1 - p_i) h_i(x) \right\}^d \right].$$

Für $d = 1$ (y wurde beobachtet, $y > 0$) folgt

$$f_i(x, 1) = \left[p_i^0 \left\{ (1 - p_i) h_i(x) \right\}^1 \right] = h_i(x) - p_i h_i(x)$$

und für $d = 0$ (y fehlt oder $y = 0$) entsprechend

$$f_i(x, 0) = \left[p_i^1 \left\{(1 - p_i)h_i(x)\right\}^0\right] = h_i(x) - p_i.$$

Diese Wahrscheinlichkeitsfunktion entspricht der bedingten Verteilung von x, dem stetigen Response, multipliziert mit der (binomialen) Wahrscheinlichkeit von d in der i-ten Population. p_i ist dabei der Anteil der Nullen und $h_i(x)$ die Verteilung von x in der i-ten Gruppe

Für die Nullhypothese gilt

$$H_0 : (p_1 = p_2) \cap (\mu_1 = \mu_2).$$

μ entspricht dem Lageparameter von $h_i(x)$, p dem jeweiligen Nullanteil der Gruppen. Somit basiert der Test wiederum auf zwei weiteren Tests: einem Test auf Gleichheit der Anteile der Nullen, $(p_1 = p_2)$, und einem Test auf die Gleichheit der Verteilungen der Werte, welche ungleich Null sind $(\mu_1 = \mu_2)$. Falls der Anteil der Nullen und die Mittelwerte der Ausprägungen, welche größer Null sind, in den Untergruppen verschieden sind, besteht die Möglichkeit, dass die Mittelwerte der Übergruppen dennoch gleich sind. Der Grund dafür ist, dass ein größerer Anteil Nullen einen hohen Mittelwert stark verringert (im Folgenden als dissonant bezeichnet) und umgekehrt (konsonant). Für den stetigen Teil wird eine spezielle Verteilung angenommen, wie zum Beispiel die Log-Normalverteilung oder die log-Gammaverteilung.

In dieser Arbeit werden die Two-Part-Models mit den bekannten Zwei-Stichproben-Tests gemacht und es folgt

$$X^2 = B^2 + T^2, \quad X^2 \sim \chi^2(2).$$

B ist dabei der Wert der Teststatistik des Binomialtests, T entweder der Wert der Teststatistik des t-Tests oder des Wilcoxon-Tests. Für B und T gilt, dass sie unter der Annahme unabhängiger Fehler der binomialen und stetigen Teile der Verteilung selbst auch unabhängig sind. Diese Tests mit zwei Freiheitsgraden sind besser geeignet als die einfachen Tests, wie nur der t-Test oder nur der Wilcoxon-Test, falls der größere Anteil an Nullen

in der Gruppe mit dem größeren Mittelwert ist. Ist dies nicht der Fall, so beeinflusst der Unterschied des Nullenanteils den Unterschied zwischen den Mittelwerten und vor allem der Wilcoxon-Test ist besser geeignet. Der cut-off Wert innerhalb des dichotomen Anteils basiert auf *a priori* Überlegungen und wird nicht aus den Daten generiert.

3 Maximal selektierte χ^2-Statistiken für ordinale Variablen

Die Beschreibung folgender Methode basiert auf [Boulesteix (2006)]. Vor allem in der Medizin steht häufig die Problemstellung im Vordergrund, dass man eine Abhängigkeit zwischen einer binären Variable Y und einer mindestens ordinal skalierten Variable X untersuchen möchte. Bei nominal skaliertem X könnte man den exakten Test nach Fisher rechnen oder, falls der Stichprobenumfang groß genug ist, auch einen asymptotischen χ^2-Test. Bei stetigem X eignen sich Tests wie der t-Test oder der Wilcoxon-Rangsummen-Test. Bei einem mindestens ordinal skalierten, aber nicht stetigem X, ist dies komplizierter. Die Verteilung der maximal selektierten χ^2-Statistik ist unter der Nullhypothese, dass X und Y unabhängig sind, verschieden von der bekannten χ^2-Verteilung. Die Abhängigkeit wird mit Hilfe eines *Cutpoints* getestet. Die maximal selektierte χ^2-Statistik entspricht der maximalen χ^2-Statistik über alle diese *Cutpoints*. Diese im Folgenden resultierende, unter der Nullhypothese geltende Verteilung der maximal selektierten χ^2-Statistik, kann ebenfalls als Messmethode der Abhängigkeit zwischen X und Y benutzt werden.

Datengrundlage ist die Stichprobe $(x_i, y_i)_{i=1,\ldots,N}$ mit N unabhängig und identisch verteilten Realisationen von X und Y. X nimmt dabei K verschiedene Level, $a_1, \ldots, a_k \in R$, an. Es gilt $2 \leq K \leq N, a_1 < \ldots < a_k$. Y besitzt die Level $Y = 1$ und $Y = 2$. Eine Möglichkeit, die Abhängigkeit zwischen X und Y zu messen, ist X in binäre Variablen $x^{(k)}$, $k = 1, \ldots, K-1$, zu

transformieren, wobei gilt:

$$x^{(k)} = 0, \quad X \leq a_k$$
$$x^{(k)} = 1, \quad \text{sonst}$$

Die resultierende Verteilung der maximal selektierten χ^2-Statistik ist abhängig von N_1 und N_2 und $m_1, \ldots, m_k$, mit

$$m_k = \sum_{i=1}^{n} I(x_i = a_k), \quad k = 1, \ldots, K,$$

wobei $I(x)$ die Indikatorfunktion ist. Man betrachte folgende 2 x 2 Kontingenztafel für $k = 1, \ldots, K-1$:

	$X \leq a_k$	$X > a_k$	Σ
$Y = 1$	$n_{1,\leq a_k}$	$n_{1,>a_k}$	N_1
$Y = 2$	$n_{2,\leq a_k}$	$n_{2,>a_k}$	N_2
Σ	$n_{.,\leq a_k} = \sum_{j=1}^{k} m_j$	$n_{.,>a_k} = \sum_{j=k+1}^{K} m_j$	N

N_1 und N_2 bezeichnen die Anzahl der Realisationen mit $y_i = 1$ und $y_i = 2$. Die entsprechende χ^2-Statistik ist

$$\chi^2_k = \frac{N(n_{1,\leq a_k} n_{2,>a_k} - n_{1,>a_k} n_{2,\leq a_k})^2}{N_1 N_2 n_{.,\leq a_k} n_{.,>a_k}}.$$

Weiter ist die maximal selektierte χ^2-Statistik definiert als

$$\chi^2_{max} = \max_{k=1,\ldots,K-1} \chi^2_k.$$

Bestimmt man ein frei wählbares d, so gilt, dass $\chi^2_{max} \leq d$ g.d.w. alle Punkte mit den Koordinaten $(n_{1,\leq a_k}, n_{2,\leq a_k})$ für $k = 1, \ldots, K-1$ auf oder über der Funktion

$$\text{lower}_d(x) = \frac{N_2 x}{N} - \frac{N_1 N_2 \sqrt{d}}{N} \sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

oder auf oder unter der Funktion

$$\text{upper}_d(x) = \frac{N_2 x}{N} + \frac{N_1 N_2 \sqrt{d}}{N} \sqrt{\frac{x}{N}\left(1 - \frac{x}{N}\right)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

liegen. Mit Hilfe von Kombinatorik kommt man zu dem Ergebnis, dass

$$P_{H_0}(\chi^2_{max} > d) = \begin{pmatrix} N \\ N_2 \end{pmatrix}^{-1} \sum_{s=1}^{q} \begin{pmatrix} N - i_s \\ N_2 - j_s \end{pmatrix} b_s,$$

$$b_1 = \begin{pmatrix} i_1 \\ j_1 \end{pmatrix},$$
$$b_s = \begin{pmatrix} i_s \\ j_s \end{pmatrix} - \sum_{r=1}^{s-1} \begin{pmatrix} i_s - i_r \\ j_s - j_r \end{pmatrix} b_r, \qquad s = 2, \ldots, q,$$

wobei b_s der Anzahl der Pfade in $\mathcal{P}_s$ entspricht und $\mathcal{P}_s$ der Menge der Pfade von $(0,0)$ bis B_s, die nicht durch $B_1, \ldots, B_{s-1}$ gehen. $B_1, \ldots, B_q$ haben die Koordinaten $(i_1, j_1), \ldots, (i_q, j_q)$ und $i = n_{\cdot,\leq a_k}$, $upper_d(i) < j \leq min(N_2, i)$ oder $max(0, i - N_1) \leq j < lower_d(i)$. Daraus ergibt sich die Verteilungsfunktion

$$F(d) = 1 - \begin{pmatrix} N \\ N_2 \end{pmatrix}^{-1} \sum_{s=1}^{q} \begin{pmatrix} N - i_s \\ N_2 - j_s \end{pmatrix} b_s.$$

Will man nun die Abhängigkeit zwischen X und Y messen, so nutzt man $F(\chi^2_{max})$. Hier testet man die Nullhypothese, dass X und Y unabhängig sind. Die Verteilungsfunktion nimmt Werte im Intervall $[0, 1]$ an. Je größer der Wert von $F(\chi^2_{max})$, desto höher ist auch der Zusammenhang zwischen X und Y, und desto kleiner ist der p-Wert. Die Verteilungsfunktion ist ein gutes Maß für den Zusammenhang, da man den p-Wert anhand von $1 - F(\chi^2_{max})$ berechnen kann.

4 Simulationsdesign

Wie in der Einleitung bereits erwähnt, soll das Verhalten der verschiedenen Tests in unterschiedlichen Situationen miteinander verglichen werden. Dazu wurden sechs verschiedene Simulationen durchgeführt. Aus jeder Simulation resultiert eine andere Verteilung der Variable X. X besteht aus zwei Stichproben, n_0 und n_1, deren Anteil an Nullen p_0 bzw. p_1 entspricht. Die zweite Variable, Y, ist binär und besteht aus den Werten 0 und 1. Der Wert 0 ist der ersten Gruppe zugeordnet, kommt somit n_0-mal vor, der Wert 1 gehört zur zweiten Gruppe und hat eine Häufigkeit von n_1. Die verwendeten Testbefehle innerhalb der Simulationen sind `t.test()`, `wilcox.test()`, `ks.test()`, `chisq.test()` und `maxsel.test()`. Für letzteren wird auf Funktionen aus dem Paket `exactmaxsel`, welches mit dem Paket `combinat` läuft, zurückgegriffen. Weiter wurde die Funktion `pchisq()` benutzt, um den p-Wert, der sich aus der Summe der Teststatistiken von t- und χ^2- und Wilcoxon- und χ^2-Test ergibt, zu erhalten. Um ein verlässliches Ergebnis zu erhalten, betrug die Anzahl der Iterationen 5000. Das bedeutet, dass 5000 mal Daten generiert wurden, auf dessen Grundlage die Tests gerechnet wurden. Für jede Iteration wurde der entsprechende Seed gesetzt. Die Unterschiede zwischen den Simulationen werden in den nächsten Unterkapiteln aufgeführt.

4.1 Simulation 1

Die erste Simulation generiert X bestehend aus Realisationen einer exponentialverteilten Zufallsvariablen und Nullen. Die ersten n_0 Werte sind unter dem Parameter λ_0 verteilt, die zweite Stichprobe, aus n_1 Werten bestehend, unter λ_1. Die Simulation wurde mit 15 verschiedenen Einstellungen wiederholt. Dabei wurden jeweils alle oben genannten Tests gerechnet. Die Settings pro Durchgang sind Tabelle 1 zu entnehmen. Die Simulation wurde pro Durchgang dreimal durchgeführt. Die Stichprobenumfänge waren $n_0 = n_1 = 50, n = 100$, $n_0 = n_1 = 20, n = 40$ und $n_0 = n_1 = 10, n = 20$.

Durchgang	p_0	p_1	λ_0	λ_1	
1	0.5	0.5	1	1	Nullhypothese!
2	0.5	0.5	1	2	
3	0.5	0.5	1	3	
4	0.5	0.5	1	5	
5	0.5	0.7	1	1	
6	0.5	0.8	1	1	
7	0.5	0.9	1	1	
8	0.5	0.7	1	2	
9	0.5	0.8	1	2	
10	0.5	0.7	1	3	
11	0.5	0.8	1	3	
12	0.5	0.7	2	1	
13	0.5	0.8	2	1	
14	0.5	0.7	3	1	
15	0.5	0.8	3	1	

Tabelle 1: Settings pro Durchgang für Simulation 1,3,4,5,6.

4.2 Simulation 2

Mit Hilfe von Simulation 2 sollen gleichverteilte Werte für X erzeugt werden. Diese sind in der ersten Teilgruppe unter dem Parameter μ_0 und in der zweiten Teilgruppe unter dem Parameter μ_1 verteilt. Hinzu kommt der Anteil der Nullen, p_0 und p_1, und zusätzlich kleine Werte in Form von Einsen und Zweien. Dies wurde erreicht, indem zufällig ausgewählte Werte, entsprechend p_0 und p_1, der beiden Stichproben durch Werte einer poissonverteilten Zufallsvariable mit Parameter $\lambda = 1$ ersetzt wurden. Y ist binär. Die Settings pro Durchgang sind Tabelle 2 zu entnehmen. Diese Simulation wurde für $n_0 = n_1 = 20$ durchgeführt und es wurden die Tests Wilcoxon-Rangsummen-Test, t-Test, Kolmogorov-Smirnov-Test und die Maxsel-Methode angewandt.

4.3 Simulation 3, 4, 5 und 6

Alle Simulationen wurden für $n_0 = n_1 = 50$ ($n = 100$) und $n_0 = n_1 = 25$ ($n = 50$) gerechnet. Der Anteil der Nullen liegt auch hier bei p_0 und p_1 und die Settings sind in Tabelle 1 abzulesen. Y ist immer binär. Das X in Simulation

Durchgang	p_0	p_1	μ_0	μ_1	
1	0.5	0.5	50	50	Nullhypothese!
2	0.5	0.5	50	30	
3	0.5	0.5	50	20	
4	0.5	0.5	50	10	
5	0.5	0.7	50	50	
6	0.5	0.8	50	50	
7	0.5	0.9	50	50	
8	0.5	0.7	50	30	
9	0.5	0.8	50	30	
10	0.5	0.7	50	20	
11	0.5	0.8	50	20	
12	0.5	0.7	30	50	
13	0.5	0.8	30	50	
14	0.5	0.7	20	50	
15	0.5	0.8	20	50	

Tabelle 2: Settings pro Durchgang für Simulation 2.

3 besteht aus logarithmisch normalverteilten Werten und Nullen, die echt positiven Werte in Simulation 4 sind normalverteilt. Simulation 5 entspricht Simulation 3 und Simulation 6 entspricht Simulation 4 mit dem Unterschied, dass es in X nicht nur die Nullen als kleine Werte gibt, sondern auch Einsen und Zweien. Diese wurden wie in Simulation 2 erzeugt. Hier wurde nur für $n_0 = n_1 = 50$ simuliert. Bei allen vier Simulationen gilt Var(X)=1.

5 Simulationsergebnisse

Um die Ergebnisse der im vorangegangen Abschnitt beschriebenen Simulationen besser beurteilen zu können, wurden unter anderem Boxplots der p-Werte erstellt. So ist schnell erkennbar, ob sich die p-Werte eher im Bereich der 0 ansammeln oder hoch sind. So kann man erste (grobe) Schlüsse über die Power der Tests unter den verschiedenen Bedingungen ziehen. Die Ergebnisse hierzu befinden sich im Anhang. Im weiteren Verlauf der Analyse wurden die Ablehnungsanteile samt Konfidenzintervall je Test und Durchlauf berechnet. Dazu wurden p-Werte < 0.05 als signifikant eingestuft. Mit die-

sen Berechnungen erhält man gleichzeitig einen Überblick über die Power der Tests, vorausgesetzt die Alternativhypothese (ein Unterschied in den Verteilungen) liegt vor. Da ein Vergleich der Ergebnisse anhand des Mittelwertes bei p-Werten nicht geeignet ist, sind die Mediane in Betracht gezogen worden. Um diese auf signifikante Unterschiede unter den einzelnen Tests zu prüfen, wurde der Wilcoxon-Rangsummen-Test gerechnet.

5.1 Simulation 1

Wie bereits erwähnt, wurde Simulation 1 für drei verschiedene Gruppengrößen durchgeführt. In den Durchgängen 2-15 lag ein Unterschied in den Anteilen der Nullen, der Mittelwerte oder beidem vor. Bei Betrachtung und Vergleichen, welcher Test am häufigsten den höchsten Ablehnungsanteil und somit auch die höchste Power besitzt, stellt sich heraus, dass sich die Ergebnisse mit den Gruppengrößen verändern (vgl. Abb. 1). Hier wird der erste Durchlauf außer Acht gelassen, da in diesem die Nullhypothese, d.h. identische Verteilungen, vorliegt, und somit das Ergebnis des Ablehnungsanteils nicht der Power entspricht und weiter auch nicht mit denen der übrigen Durchläufe vergleichbar ist.

Für $n_0 = n_1 = 50$ erreichen der Wilcoxon-Test und der t-Test im Two-Part-Modell (BT) am häufigsten den höchsten Ablehnungsanteil (jeweils 8 von 14), gefolgt von der Maxsel-Methode, welche in 7 von 14 Durchgängen eine der Methoden mit dem größten Ablehnungsanteil ist. Platz 3 teilen sich der Fisher -, Kolmogorov-Smirnov- und Wilcoxon-Test im Two-Part-Modell (BW) mit der höchsten Power in 6 von 14 Durchgängen. Schwächster Test ist der t-Test, welcher nur in 3 von 14 Fällen den höchsten Ablehnungsanteil im Vergleich mit den anderen Tests erreicht.

Geht man in der Betrachtung weiter ins Detail und überprüft, welche Voraussetzungen jeweils vorliegen und wie hoch der tatsächliche Ablehnungsanteil ist, so erkennt man, dass der des Fisher-Tests entweder 0 (95% KI, 0-0) oder 1 (95%KI, 1-1) annimmt. Die gleiche Situation liegt beim Wilcoxon-Test im Two-Part-Modell (BW) vor. Auch die anderen Tests erreichen, abgesehen vom t-Test, teilweise einen Ablehnungsanteil von 1. Auffällig ist, dass, so-

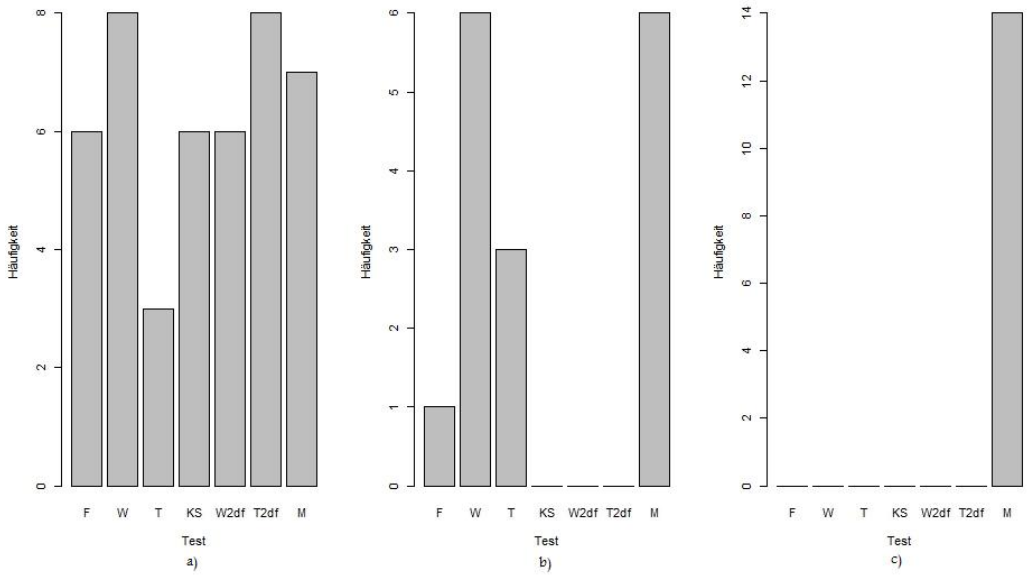

Abbildung 1: Häufigkeiten, wie oft den Tests der maximale Ablehnungsanteil zugeordnet wird (Durchgang 2-15) für n=100 (a), n=40 (b), n=10 (c), (Simulation 1), falls mehreren Tests der maximale Anteil zugeteilt wird, erhöht sich die Häufigkeit bei jedem entsprechend um 1.

bald nur einer dieser Tests diesen höchstmöglichen Anteil erreicht, auch alle anderen Tests (bis auf den t-Test) die Nullhypothese in jeder der 5000 Iterationen ablehnen und sich somit für die Alternativhypothese entscheiden. In diesen Fällen beträgt die Differenz der Anteile der Nullen mindestens 0.3. Der t-Test, welcher die Mittelwerte untersucht, lehnt die Nullhypothese besonders oft ab, falls sich die Mittelwerte unterscheiden, die Anteile der Nullen in den beiden Gruppen jedoch identisch sind (hier $p_0 = p_1 = 0.5$). Der Ablehnungsanteil erhöht sich mit zunehmender Differenz zwischen den Mittelwerten (0.416, 95% KI, 0.402-0.429; 0.864, 95% KI, 0.855-0.874; 0.994, 95% KI, 0.991-0.996). Dass die Daten nicht normalverteilt sind scheint kein größeres Problem darzustellen, da die Approximation einer solchen durch $n > 30$ gegeben ist. Mit Blick auf die letzte Spalte der Tabelle 3 erkennt man jedoch schnell, dass der t-Test, abgesehen von den eben beschriebenen drei Durchgängen, meist weit hinter den anderen liegt. Der größte Unterschied

zu den übrigen Tests liegt in den dissonanten Fällen vor, d.h. wenn in der Gruppe mit größerem Mittelwert auch der höhere Nullanteil vorhanden ist, was dazu führt, dass der Gesamtmittelwert geringer wird.

Der Kolmogorov-Smirnov-Test lehnt die Nullhypothese am öftesten in Situationen ab, in denen die Differenz der Nullanteile größer ist. Diese Differenz hat größeren Einfluss auf den Ablehnungsanteil als Unterschiede in den Mittelwerten innerhalb der Gruppen ohne die Nullen. Die Ursache ist darin zu vermuten, dass große Unterschiede in den Nullanteilen auch große Unterschiede in den Verteilungen bedeuten, welche der Kolmogorov-Smirnov-Test ursprünglich untersucht.

Die Maxsel-Methode hat zwar nicht immer den höchsten Ablehnungsanteil, ist jedoch, bis auf ein paar Ausnahmen, konstant gut auf die verschiedenen Gegebenheiten eingegangen. Am meisten Probleme, einen Unterschied zwischen den Gruppen zu erkennen, gab es bei geringem Unterschied in den Mittelwerten und entgegengesetztem geringen Unterschied in den Nullanteilen (0.063, 95% KI, 0.056-0.069) oder wenn es nur einen geringen Unterschied in den Nullanteilen gibt, jedoch keinen zwischen den Mittelwerten (0.192, 95% KI, 0.181-0.203). Hier ist jedoch dringend anzumerken, dass bei diesen Gegebenheiten alle anderen Tests einen noch geringeren Ablehnungsanteil aufweisen. Besonders der Kolmogorov-Smirnov-Test hat hier aufgrund des zuvor angemerkten Zusammenhangs zwischen Nullanteil und Verteilung eine schwache Power.

Ein Vergleich zwischen den Two-Part-Modellen liefert als Ergebnis, dass der BT genau so gut geeignet zu sein scheint wie der BW, häufig sogar besser. Bei einem Unterschied der Nullanteile von 0.2 und einem Unterschied zwischen λ_0 und λ_1 von 2, sowohl im konsonantischen als auch im dissonanten Fall, führt der BT das Feld an. Im konsonantischen ist der Ablehnungsanteil des Tests, wie auch bei allen anderen, jedoch wesentlich höher (0.998, 95% KI 0.997-0.993) als im dissonanten (0.249, 95% KI 0.237-0.261). Der Wilcoxon-Test scheint am besten geeignet, um einen Unterschied aufzudecken, falls der Unterschied in den Nullanteilen zwar vorhanden aber gering ist, und auch der Unterschied zwischen den Mittelwerten der echt positiven Werte eher klein ist (0.961, 95% KI, 0.956-0.967) oder es keinen gibt (0.997, 95% KI,

0.996-0.999).

Als nächstes wird die Situation im Fall $n_0 = n_1 = 20$, und somit $n = 40$, betrachtet. Die sich ergebenden Ablehnungsanteile der einzelnen Tests sind in Tabelle 4 abzulesen. Wie auch zuvor, liegt der Ablehnungsanteil beim Fisher-Test entweder bei genau 0 oder genau 1. Letzteres kommt nur einmal vor. Hier ist ein Unterschied von 0.4 zwischen den Nullanteilen bei gleichen Gruppenmittelwerten nötig. Der Wilcoxon-Test erreicht nur dreimal einen Ablehnungsanteil > 0.5. Die Situation ist jedes Mal konsonantisch, mit $p_0 - p_1 = -0.3$ bzw. -0.4 und $\lambda_0 - \lambda_1 = -1$ bzw. 0 (0.744, 95% KI, 0.732,0.756; 0.903, 95% KI, 0.895-0.911 und 1, 95% KI 1-1). In diesen Fällen ist der Ablehnungsanteil des Wilcoxon-Tests im Vergleich mit denen der anderen der höchste. Er ist hauptsächlich am geringsten, wenn die Nullanteile gleich sind, unabhängig von der Differenz zwischen den Mittelwerten. Der Grund hierfür liegt bei den ursprünglichen Absichten des Tests. Der Wilcoxon soll testen, ob ein Unterschied in den Medianen der beiden Gruppen vorliegt. Da die Nullen in beiden Gruppen bereits jeweils 50% ausmachen, und der Median den Wert angibt, unterhalb welchem 50% der Daten liegen, wird dieser, da die Anzahl der Werte je Gruppe eine gerade ist, aus dem Mittel der 0 und dem Minimum der jeweiligen Gruppe gebildet. Diese beiden Minima unterscheiden sich ohnehin nicht groß voneinander. Durch die Hinzunahme der 0 und Bilden des Mittelwerts gleichen sie sich weiter aneinander an. Auch die anderen Tests haben in diesen Fällen eine eher geringe Power. Den höchsten Ablehnungsanteil hat noch die Maxsel-Methode bei $\lambda_0 - \lambda_1 = -4$ (0.629, 95% KI 0.615-0.642). Der Kolmogorov-Smirnov-Test lehnt die Nullhypothese bei diesen Gruppengrößen nur sehr selten ab. Der maximale Ablehnungsanteil liegt hier bei 0.172 (95% KI, 0.162-0.182). Der BW erkennt zu keiner Zeit, dass die Alternativhypothese vorliegt. Auch die Power des t-Tests und des BTs nimmt deutlich ab, was wohl an der nicht mehr gegebenen Approximation der Normalverteilung liegt. Einen wirklich hohen Ablehnungsanteil hat bei diesen Gruppengrößen keiner der Tests mehr. Am ehesten scheint in den meisten Situationen noch die Maxsel-Methode zu empfehlen zu sein, die, im Vergleich mit den anderen Tests, den höchsten Ablehnungsanteil besitzt, falls die Nullanteile identisch sind oder sich, in dissonanten Fällen, nur gering

F	W	T	KS	BW	BT	M	
0	0	0.0138	6e-04	0	0.0014	0.0188	Nullhypothese!
0	0	0.4158	0.0544	0	0.1806	0.359	(F W BW) KS BT M T
0	0.0016	0.8644	0.278	0	0.6338	0.7964	(F BW) W KS BT M T
0	0.0422	0.9936	0.7528	0	0.9596	0.9876	(F BW) W KS BT M T
0	0.4724	0.1976	0.0618	0	0.3472	0.1916	(F BW) KS M T BT W
1	1	0.5766	1	1	1	1	T (F W KS BW BT M)
1	1	0.914	1	1	1	1	T (F W KS BW BT M)
0	0.9614	0.871	0.3766	0	0.957	0.766	(F BW) KS M T BT W
1	1	0.9744	1	1	1	1	T (F W KS BW BT M)
0	0.9974	0.9898	0.6874	0	0.9982	0.9602	(F BW) KS M T W BT
1	1	0.9986	1	1	1	1	T (F W KS BW BT M)
0	0.04	0.008	0.0056	0	0.0384	0.0626	(F BW) KS T BT W M
1	1	0.0626	1	1	1	1	T (F W KS BW BT M)
0	0.0034	0.0746	0.0022	0	0.2486	0.2384	(F BW) KS W T M BT
1	1	0.0084	1	1	1	1	T (F W KS BW BT M)

Tabelle 3: Ablehnungsanteile für Simulation 1 je Durchlauf (1-15) für $n_0 = n_1 = 50$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

F	W	T	KS	BW	BT	M	
0	0	0.0138	6e-04	0	0.0014	0.0188	Nullhypothese!
0	0	0.0668	0.0038	0	0.0156	0.104	(F W BW) KS BT T M
0	0	0.2228	0.0194	0	0.061	0.2936	(F W BW) KS BT T M
0	0	0.5426	0.0944	0	0.2348	0.6288	(F W BW) KS BT T M
0	0.0032	0.051	0.0022	0	0.0272	0.0452	(F BW) KS W BT M T
0	0.3654	0.177	0.0136	0	0.3002	0.222	(F BW) KS T M BT W
1	1	0.4942	0.172	0	0.9786	1	BW KS T BT (F W M)
0	0.0298	0.281	0.0212	0	0.1844	0.2472	(F BW) KS W BT M T
0	0.7438	0.503	0.064	0	0.6914	0.5216	(F BW) KS T M BT W
0	0.091	0.515	0.061	0	0.3666	0.4714	(F BW) KS W BT M T
0	0.9028	0.697	0.137	0	0.8646	0.714	(F BW) KS T M BT W
0	4e-04	0.0026	0	0	0.0012	0.006	(F KS BW) W BT T M
0	0.102	0.0294	0.0014	0	0.0612	0.0706	(F BW) KS T BT M W
0	0	0.0014	0	0	2e-04	0.0114	(F W KS BW) BT T M
0	0.0376	0.0108	6e-04	0	0.0194	0.0338	(F BW) KS T BT M W

Tabelle 4: Ablehnungsanteile für Simulation 1 je Durchlauf (1-15) für $n_0 = n_1 = 20$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

unterscheiden.

Im letzten Teil der Simulation wurden die Gruppengrößen noch einmal auf $n_0 = n_1 = 10$ reduziert. Die Ablehnungsanteile sind in Tabelle 5 zu finden. Hier fallen sofort die sehr geringen Ablehnungsanteile auf, welche nur in wenigen Fällen > 0.1 sind. Die Anteile, welche größer als 0.1 sind, stammen von der Maxsel-Methode, mit einer Ausnahme. Der Wilcoxon-Test, welcher ansonsten zu keiner Zeit einen Unterschied erkennt und der Ablehnungsanteil somit bei 0 liegt, erreicht für $p_0 - p_1 = -0.4$ einen Anteil von 0.162 (95% KI, 0.152-0.173). Auch die anderen Tests sind in diesem Fall, verglichen mit ihren anderen Ergebnissen, am stärksten. Für sich genommen sind jedoch auch diese Ergebnisse sehr schwach. Für den Fall, dass die Nullhypothese vorliegt, erkennen die Tests dies zu sehr niedrigen Signifikanzniveaus. Der höchste p-Wert stammt hier von der Maxsel-Methode und liegt bei 0.009. Diese Werte sind im Vergleich mit den Ergebnissen der Simulationen mit größeren Stichproben ziemlich klein.

F	W	T	KS	BW	BT	M	
0	0	8e-04	0	0	4e-04	0.0092	Nullhypothese!
0	0	0.004	0	0	0.0016	0.0292	(F W KS BW) BT T M
0	0	0.012	0	0	0.0034	0.068	(F W KS BW) BT T M
0	0	0.0334	0	0	0.0118	0.1654	(F W KS BW) BT T M
0	0	0.0048	0	0	0.0018	0.0194	(F W KS BW) BT T M
0	0	0.0178	0	0	0.016	0.0468	(F W KS BW) BT T M
0	0.1624	0.0738	0	0	0.2072	0.3234	(F KS BW) T W BT M
0	0	0.0208	0	0	0.0064	0.068	(F W KS BW) BT T M
0	0	0.046	0	0	0.042	0.1202	(F W KS BW) BT T M
0	0	0.042	0	0	0.0144	0.1274	(F W KS BW) BT T M
0	0	0.0788	0	0	0.0734	0.196	(F W KS BW) BT T M
0	0	0.001	0	0	2e-04	0.0044	(F W KS BW) BT T M
0	0	0.0054	0	0	0.0044	0.0172	(F W KS BW) BT T M
0	0	4e-04	0	0	0	8e-04	(F W KS BW BT) T M
0	0	0.0022	0	0	0.002	0.0078	(F W KS BW) BT T M

Tabelle 5: Ablehnungsanteile für Simulation 1 je Durchlauf (1-15) für $n_0 = n_1 = 10$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

5.2 Simulation 2

Wie schon bei Simulation 1 wurde zunächst für einen ersten Überblick ein Balkendiagramm bezüglich der Häufigkeiten der maximalen Ablehnungsanteile je Durchgang der einzelnen Tests erstellt (vgl. Abb. 2). Wichtig ist, sich bei der Betrachtung der Ergebnisse in Erinnerung zu rufen, dass hier nicht nur Nullen den angegebenen Anteil ausmachen, sondern auch kleine Zahlen wie 1,2 u.s.w. Einfachheitshalber wird dieser Anteil dennoch als Nullanteil bezeichnet! Auch hier wurden, aus demselben Grund wie bei vorheriger Simulation, nur die Durchläufe 2-15 betrachtet. Diese Simulation wurde mit dem Wilcoxon-, dem t-, dem Kolmogorov-Smirnov-Test und der Maxsel-Methode gemacht und für $n_0 = n_1 = 20$ durchgeführt.

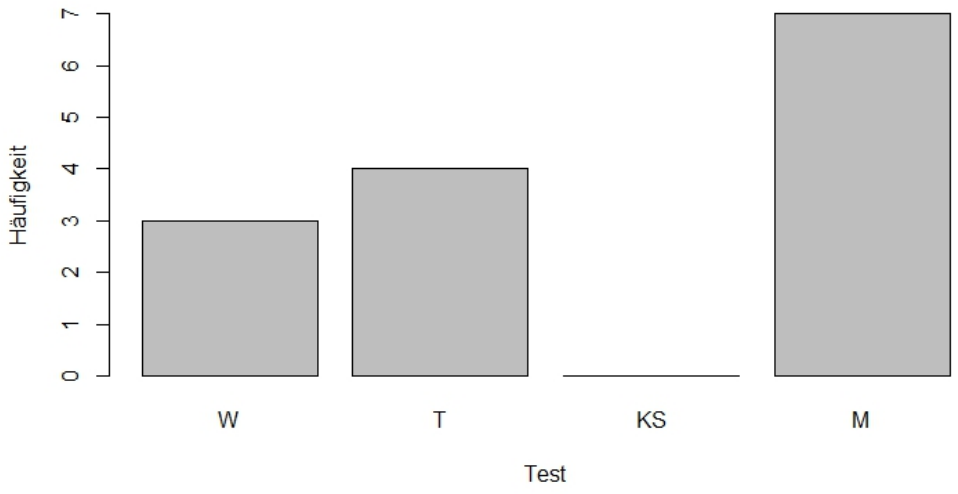

Abbildung 2: Häufigkeiten, wie oft den Tests der maximale Ablehnungsanteil zugeordnet wird für n=40, (Simulation 2), falls mehreren Tests der maximale Anteil zugeteilt wird, erhöht sich die Häufigkeit bei jedem entsprechend um 1.

Die Maxsel-Methode hat hier eindeutig am häufigsten, in 50% der Fälle, die höchste Power. Es folgt der t-Test in 4 und der Wilcoxon-Test in 3 von 14 Durchläufen. Der Kolmogorov-Smirnov-Test hingegen ist kein Mal der stärkste Test.

Einen detaillierteren Überblick schafft Tabelle 6. Liegt die Nullhypothese vor, d.h. sind die Verteilungen in den beiden Gruppen identisch, so erkennen der Wilcoxon- und der Kolmogorov-Smirnov-Test dies am häufigsten. Sie behalten die Nullhypothese immer bei. Auch der t-Test lehnt H_0 nur in den wenigsten Fällen ab. Die Maxsel-Methode lehnt die Nullhypothese in etwas mehr als 1% der Fälle ab. Dieser Wert liegt jedoch ebenso weit unter dem vorgegebenen Signifikanzniveau von 0.05. Somit scheinen alle vier Tests verlässlich zu sein, sollten sie für eine identische Verteilung entscheiden.

Für die weiteren Durchläufe fällt auf, dass, bis auf die vier letzten und den sechsten, Maxsel-Methode und t-Test aufeinander folgen. In den Durchläufen, bei denen sich die Gruppen nur in den Mittelwerten unterscheiden, ist die Maxsel-Methode stärker in der Detektierung eines Unterschieds zwischen den Verteilungen. Die Differenz in der Power beider Tests nimmt jedoch ab, je größer der Unterschied zwischen den Mittelwerten des echt positiven Teils der Gruppen wird. So beträgt sie bei der Maxsel-Methode in Durchlauf 3 bei einem Nullanteil von je 0.5 und einer Differenz zwischen den Mittelwerten von $50 - 20 = 30$ bereits 0.712 (95% KI, 0.67-0.725), beim t-Test jedoch erst 0.389 (95% KI, 0.376-0.403). In Durchgang 4 jedoch verbessert sich die Maxsel-Methode um ca. 38% auf 0.980 (95% KI, 0.976-0.984), der t-Test hingegen verbessert sich um ganze 142% auf 0.941 (95% KI, 0.935-0.948). Kolmogorov-Smirnov und Wilcoxon haben in diesen Fällen eine Power < 0.1, werden jedoch auch mit Zunahme der Mittelwertdifferenz stärker. Dies ist allerdings nur relativ zu sehen, denn der Wilcoxon erreicht in Durchlauf 4 nur eine Power von 0.048 (95% KI, 0.042-0.054), der Kolmogorov-Smirnov kommt immerhin auf 0.439 (95% KI, 0.425-0452). In den Fällen, in denen sich die Gruppen in den Nullanteilen unterscheiden und die Mittelwerte der echt positiven Werte der Stichproben identisch sind, haben alle vier Tests eine sehr geringe Power. Davon ausgenommen ist Durchlauf 7, bei welchem die Differenz zwischen dem Nullanteil in der ersten und dem Nullanteil der zweiten Gruppe -0.4 beträgt. Hier haben ein weiteres Mal die Maxsel-Methode mit 0.896 (95% KI, 0.887-0.904) und der t-Test mit 0.732 (95% KI, 0.72-0.744) die höchste Power. Nicht viel schlechter ist der Wilcoxon mit 0.626 (95% KI, 0.613-0.64). Nur der Kolmogorov-Smirnov liegt mit 0.409 (95% KI,

0.395-0.423) unterhalb einer Power von 0.5. Die Stärke der drei erstgenannten Tests lässt sich wie folgt begründen: Die Maxsel-Methode findet den Cutpoint mit der maximalen χ^2-Statistik, so dass der höchstmögliche Zusammenhang zwischen den Daten besteht. Somit sind die metrischen Daten optimal in kategorisiert. Da der Unterschied zwischen den einzelnen Gruppen sehr hoch ist, gibt es keinen Grund, dass dieser von der Methode unter den gegebenen Umständen nicht erkannt wird. Auch der t-Test, welcher auf Unterschiede in den Mittelwerten der zwei Gruppen testet, deckt den Unterschied, welcher durch den konsonantischen Effekt noch verstärkt wird, in den Verteilungen auf. Für den Wilcoxon ist der Unterschied in den Medianen nun größer und somit ersichtlicher als in zuvor genannter Situation, da der Median der zweiten Gruppe hier bei einem Anteil der Nullen von 90% definitiv einen der kleinen Werte annimmt. Der Median der ersten Gruppe dagegen bildet sich als Mittel aus dem Maximum der kleinen Werte und dem Minimum der großen Werte (welche entscheidend höher sind als die kleinen).

Bei Durchlauf 8 bis 11, den konsonantischen Fällen, dominiert der t-Test. Die Reihenfolge in der Power der Tests ist bei jedem der Durchgänge identisch, der Kolmogorov-Smirnov-Test liegt, hinter dem Wilcoxon und der Maxsel-Methode, immer an letzter Stelle. Maxsel, Kolmogorov-Smirnov und t-Test werden mit jedem Durchgang stärker, hängen somit nur von dem absoluten Unterschied zwischen den Gruppen ab. Die Power des Wilcoxon dagegen wird mit zunehmendem Unterschied der Mittelwerte des echt-positiven Teils höher, ist dabei jedoch jeweils minimal kleiner für die Durchläufe, in denen die Differenz zwischen den Nullanteilen (nur) -0.2 beträgt. Den höchsten Ablehnungsanteil hat jede Methode bei Durchlauf 11 (Differenz in Nullanteilen: -0.3, Differenz in Mittelwerten des echt-positiven Teils: 30). Der t-Test ist hier mit einer Power von 0.946 (95% KI, 0.94-0.953) ganze 92% besser als der Wilcoxon mit 0.493 (95% KI, 0.479-0.507) und sogar mehr als dreimal so stark wie der Kolmogorov-Smirnov mit nur 0.235 (95% KI, 0.223-0.247).

In den dissonanten Fällen (Durchlauf 12 bis 15), ist im Grunde genommen die Power jedes Tests ziemlich niedrig. Selten wird ein Wert von 0.1 überschritten. Der Grund dafür ist der geringe Unterschied zwischen den Gruppen. Trotz größerer Unterschiede in den Mittelwerten des echt-positiven

Teils sind die absoluten Unterschiede der Mittelwerte aufgrund des gegensätzlichen Nullanteils nur gering. So gesehen ist es unter Umständen gar nicht so falsch, dass die Tests diesen Unterschied nicht (als signifikant) detektieren, da er einfach zu gering ist.

	W	T	KS	M	
1	0	8e-04	0	0.0124	Nullhypothese!
2	0.0014	0.063	0.0076	0.2838	W KS T M
3	0.01	0.3894	0.0834	0.7124	W KS T M
4	0.0478	0.9418	0.4386	0.9798	W KS T M
5	0.0442	0.0506	0.0076	0.0526	KS W T M
6	0.2618	0.2544	0.0582	0.213	KS M T W
7	0.6264	0.732	0.409	0.8956	KS W T M
8	0.1378	0.4158	0.0316	0.381	KS W M T
9	0.4106	0.7404	0.1136	0.5472	KS W M T
10	0.2334	0.8062	0.1292	0.7642	KS W M T
11	0.4928	0.9464	0.235	0.8372	KS W M T
12	0.0094	0.0032	0.006	0.0292	T KS W M
13	0.1428	0.0364	0.0456	0.114	T KS M W
14	0.003	2e-04	0.0052	0.0848	T W KS M
15	0.0912	0.0084	0.0392	0.0844	T KS M W

Tabelle 6: Ablehnungsanteile für Simulation 2 je Durchlauf (1-15) für $n_0 = n_1 = 20$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

5.3 Simulation 3

Für einen ersten Überblick über die Ergebnissituation zunächst die Abbildung der Balkendiagramme zu den jeweiligen Durchläufen für $n_0 = n_1 = 50$ und $n_0 = n_1 = 25$ (vgl. Abb. 3).

Es ist zu erkennen, dass die Maxsel-Methode am häufigsten die höchste Power hat. Die anderen Tests unterscheiden sich, ungeachtet der tatsächlichen Power, in diesem Kriterium nicht besonders. Der t-Test zählt immer (sowohl für die größeren, als auch die kleineren Gruppengrößen) zu den Tests mit einem geringeren Ablehnungsanteil. Bei Gruppengrößen mit $n_0 = n_1 = 25$ liegen die Tests näher beieinander.

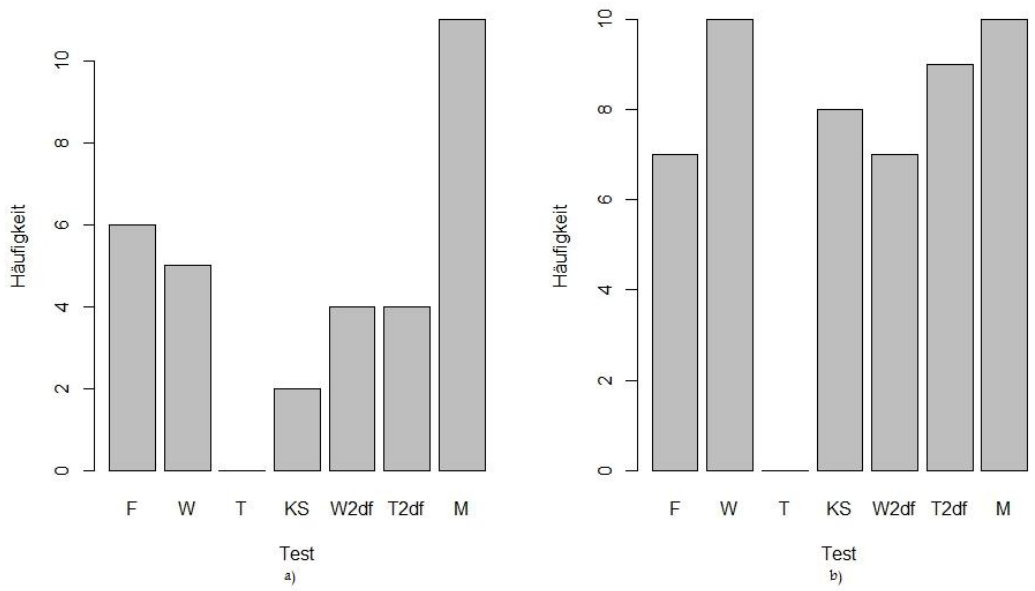

Abbildung 3: Häufigkeiten, wie oft den Tests der maximale Ablehnungsanteil zugeordnet wird für n=100 (a) und n=50 (b), (Simulation 3), falls mehreren Tests der maximale Anteil zugeteilt wird, erhöht sich die Häufigkeit bei jedem entsprechend um 1.

Für die Details betrachte man Tabelle 7 und 8. Liegt kein Unterschied in der Verteilung beider Gruppen vor, so erkennen die Tests dies sowohl bei einem Gruppenumfang von 50 als auch von 25 richtig. Gibt es den Unterschied aufgrund verschiedener Mittelwerte des echt-positiven Teils während die Nullanteile in beiden Gruppen identisch sind, so haben fast alle Tests, bei einer Gruppengröße von 50, eine Power > 0.9. Der Fisher-Test und der BW jedoch erkennen diesen Unterschied zu keiner Zeit. Ihre Power beträgt konstant 0. Einzig für den Fall, dass die Differenz zwischen den Mittelwerten $1 - 0.5 = 0.5$ beträgt, erreichen die übrigen Tests das Niveau von 0.9 nicht. Hier ist die Maxsel-Methode mit einem Ablehnungsanteil von 0.642 (95% KI, 0.629-0.655) am stärksten. Auch in den anderen beiden Fällen liegt sie vor den anderen Tests und lehnt die Nullhypothese (fast) immer korrekt ab. Ihre Power steigt über 0.999 (95% KI, 0.999-1) bis hin zu 1. Unterscheiden sich die Gruppen aufgrund der Nullanteile, jedoch nicht in den Mittelwerten

des echt-positiven Teils, so beträgt die Power jedes Tests, abgesehen vom t-Test, 1. Diese Aussage gilt für einen Unterschied in den Nullanteilen von mindestens -0.3. Bei einem Unterschied von nur -0.2 ist der Wilcoxon-Test mit einer Power von 0.498 (95% KI, 0.485-0.512) der Stärkste. Den größten Unterschied macht es für den Kolmogorov-Smirnov-Test, ob die Differenz der Nullanteile nur -0.2 oder schon -0.3 beträgt. Im ersten Fall beträgt die Power gerade einmal 0.059 (95% KI, 0.052-0.066), wohingegen sie im zweiten direkt auf 1 ansteigt. In konsonantischen Fällen sind die Tests nahezu gleich gut. Ausnahmen sind hier der Fisher-Test und BW für den Fall, dass die Differenz zwischen den Nullanteilen 0.2 beträgt und die der Mittelwerte -0.5. Hier erkennen sie zu keiner Zeit, dass ein Unterschied existiert, wohingegen die anderen Tests eine Power von (nahe) 1 erreichen. In den beiden letzten Durchläufen, in denen der Unterschied in den Mittelwerten des echt-positiven Teils $-2/3$ beträgt, ist der t-Test den anderen Tests weit unterlegen. Hier erreicht dieser nur eine Power von 0.073 (95% KI, 0.069-0.08).

Für eine Gruppengröße von jeweils 25 sind die Tests im Allgemeinen alle schwächer. Eine Ausnahme bilden hier die dissonanten Fälle, in denen alle Tests (bis auf Fisher und BW in 2 Durchläufen) eine Power nahe der 1 erreichen.

5.4 Simulation 4

Ein Blick auf die Balkendiagramme (Abb. 4) zeigt, dass bei einer Gruppengröße von $n_0 = n_1 = 50$ die Maxsel-Methode am häufigsten die höchste Power hat. Auch auf den BT ist in 50% der Fälle (mit) der meiste Verlass. Am schwächsten ist der BW, welcher niemals den höchsten Ablehnungsanteil hat. Dies gilt auch für $n_0 = n_1 = 25$. Statt der Maxsel-Methode ist hier jedoch der Fisher-Test am häufigsten der zuverlässigste.

Die Nullhypothese wird zuverlässig für beide Gruppengrößen erkannt (vgl. Tabellen 9 und 10). Bei gleichen Nullanteilen sind alle Tests bis auf Fisher und BW sehr stark, sobald die absolute Differenz zwischen den Mittelwerten größer als 2/3 ist. Bei einer kleineren Differenz von 0.5 liegt die Power bei allen Tests unter 0.5. Nur die Maxsel-Methode und der t-Test überschreiten

F	W	T	KS	BW	BT	M	
0	0	0.0086	8e-04	0	6e-04	0.0162	Nullhypothese!
0	0.0014	0.6372	0.2756	0	0.3444	0.6418	(F BW) W KS BT T M
0	0.482	0.9836	0.9886	0	0.9312	0.9994	(F BW) W BT T KS M
0	1	0.9926	1	0	0.9728	1	(F BW) BT T (W KS M)
0	0.4984	0.1796	0.059	0	0.3352	0.2054	(F BW) KS T M BT W
1	1	0.502	1	1	1	1	T (F W KS BW BT M)
1	1	0.8568	1	1	1	1	T (F W KS BW BT M)
0	8e-04	0.0238	2e-04	0	0.1252	0.1002	(F BW) KS W T M BT
1	1	0.0134	1	1	1	1	T (F W KS BW BT M)
0	0	0.6716	0.0328	0	0.9048	0.8072	(F W BW) KS T M BT
1	1	0.0704	1	1	1	1	T (F W KS BW BT M)
0	0.998	0.9106	0.8294	0	0.9724	0.9668	(F BW) KS T M BT W
1	1	0.9658	1	1	1	1	T (F W KS BW BT M)
1	1	0.0728	1	1	1	1	T (F W KS BW BT M)
1	1	0.0728	1	1	1	1	T (F W KS BW BT M)

Tabelle 7: Ablehnungsanteile für Simulation 3 je Durchlauf (1-15) für $n_0 = n_1 = 50$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

F	W	T	KS	BW	BT	M	
0	0	0.0048	8e-04	0	2e-04	0.0096	Nullhypothese!
0	0	0.2676	0.0664	0	0.0764	0.3074	(F W BW) KS BT T M
0	0	0.8294	0.6634	0	0.5736	0.937	(F W BW) BT KS T M
0	0	0.942	1	0	0.8312	1	(F W BW) BT T (KS M)
0	0.0168	0.057	0.0048	0	0.0374	0.067	(F BW) KS T M BT W
1	0.891	0.2582	0.095	0	0.63	0.4432	BW KS T M BT W F
1	1	0.5316	1	1	1	1	T (F W KS BW BT M)
0	0	0.0036	0	0	0.001	0.0494	(F W KS BW) BT T M
1	0.3018	0.0118	0.0022	0	0.0984	0.0566	BW KS T M BT W F
0	0	0.1572	0	0	0.095	0.4386	(F W KS BW) BT T M
1	0.5938	0.0846	0.0046	1	1	1	KS T W (F BW BT M)
0	0.9716	0.6516	0.565	0	0.7794	0.9146	(F BW) KS T BT M W
1	1	0.8126	0.9504	1	1	1	T KS (F W BW BT M)
0	1	0.9102	0.9898	0	0.9588	1	(F BW) T BT KS (W M)
1	1	0.927	0.9994	1	1	1	T KS (F W BW BT M)

Tabelle 8: Ablehnungsanteile für Simulation 3 je Durchlauf (1-15) für $n_0 = n_1 = 25$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

diese Marke mit 0.662 (95% KI, 0.649-0.675) und 0.696 (95% KI, 0.683-0.709). Unterscheiden sich die Verteilungen aufgrund unterschiedlicher Nullanteile, so sind die Tests eher schwach, solange die Differenz nicht 0.4 beträgt. Für diesen Fall haben alle Tests eine Power > 0.8. Eine Ausnahme bilden der

Fisher und die Maxsel-Methode, welche auch bei einem Unterschied in den Nullanteilen von 0.3 eine Power von 0.802 (95% KI, 0.791-0.813) und 0.565 (95% KI, 0.551-0.579) aufweisen.

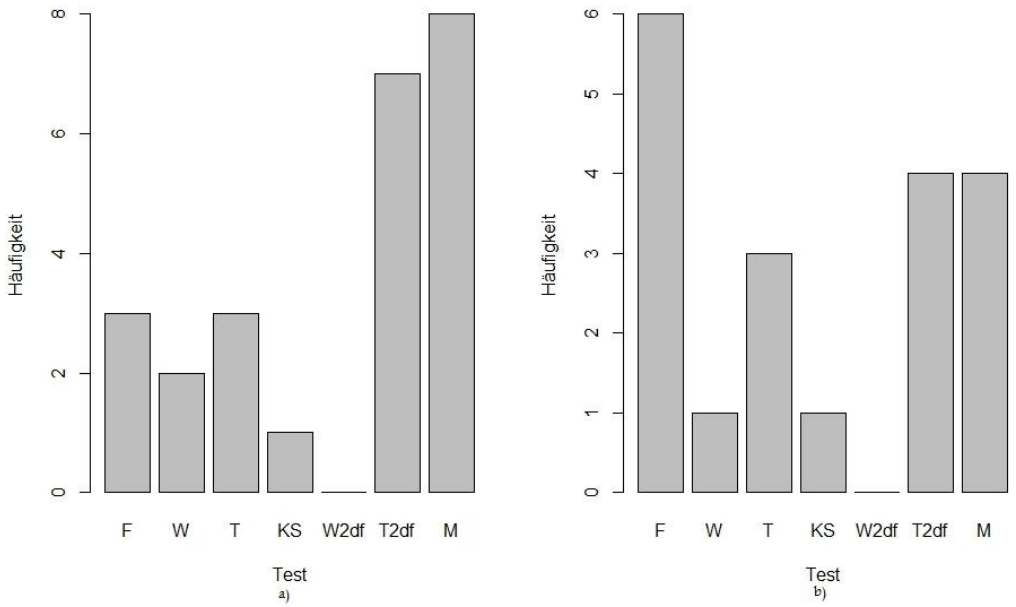

Abbildung 4: Häufigkeiten, wie oft den Tests der maximale Ablehnungsanteil zugeordnet wird für n=100 (a) und n=50 (b), (Simulation 4), falls mehreren Tests der maximale Anteil zugeteilt wird, erhöht sich die Häufigkeit bei jedem entsprechend um 1.

Bei Unterschieden sowohl in Nullanteilen als auch in den Mittelwerten im konsonantischen Sinne sind die Ergebnisse unterschiedlich. Für geringe Differenzen in Nullanteilen und Mittelwerten ist kein Test besonders gut geeignet. Die Power liegt hier bei sechs von sieben Tests unter 0.01. Nur die Maxsel-Methode erreicht einen Wert von 0.163 (95% KI, 0.153-0.173). Bis auf Fisher und Maxsel sind auch bei den übrigen Durchläufen entsprechender Einstellungen die Tests sehr schwach und ihre Power nicht nennenswert. Fisher erreicht bei Differenzen von 0.3 in den Nullanteilen immerhin eine Power von 0.829 (95% KI, 0.819-0.84) und 0.808 (95% KI, 0.797-0.819). Maxsel bei einem Unterschied in den Nullanteilen von 0.2 und einer Mittelwertdifferenz

von 2/3 eine Power von 0.835 (95% KI, 0.825-0.846). Bei identischer Mittelwertdifferenz und einem Unterschied in den Nullanteilen von 0.3 kommt die Methode nur auf 0.560 (95% KI, 0.546-0.574), womit sie jedoch einen immer noch deutlich höheren Ablehnungsanteil als die übrigen Tests hat (außer Fisher).

Für die dissonanten Fälle sticht für eine Differenz in den Mittelwerten von 0.5 nur der BW bei einem Unterschied zwischen den Nullanteilen von 0.2 heraus. Während alle anderen Tests eine Power von 0.8 überschreiten, kommt er nur auf 0.328 (95% KI, 0.315-0.341). Für einen Mittelwertunterschied von 2/3 hat vor allem der Fisher mit 0.802 (95% KI, 0.791-0.813) eine hohe Power. Die Maxsel-Methode hebt sich mit 0.565 (95% KI, 0.551-0.579) jedoch auch noch deutlich von den übrigen Tests ab.

Für die kleinere Gruppengröße von 25 Werten je Stichprobe sind die Tests, im Vergleich mit den vorherigen Simulationen, noch immer stark. Die höchste Power haben sie in den dissonanten Situationen, die geringste vor allem in den konsonantischen. Vor allem Fisher und BW erkennen bei gleichen Nullanteilen einen Unterschied in den Verteilungen aufgrund einer Differenz zwischen den Mittelwerten zu keiner Zeit, wohingegen die übrigen Tests ab einem Unterschied von 2/3 verlässlich sind.

F	W	T	KS	BW	BT	M	
2e-04	0.0084	0.0146	8e-04	0	0.005	0.0236	Nullhypothese!
4e-04	0.396	0.6962	0.2756	0	0.4712	0.6622	BW F KS W BT M T
6e-04	0.9602	0.9998	0.9886	0	0.9956	0.9994	BW F W KS BT M T
6e-04	1	1	1	0	1	1	BW F (W T KS BT M)
0.202	0.2162	0.2006	0.0476	0.034	0.327	0.2016	BW KS T M F W BT
0.8024	0.0248	0.0014	0.075	0.217	0.282	0.565	T W KS BW BT M F
0.9998	0.8808	0.893	0.955	0.9984	0.9986	1	W T KS BW BT F M
0.0078	0.0014	0.009	2e-04	2e-04	0.0018	0.1628	(KS BW) W BT F T M
0.8292	0.1546	0.0176	0.1284	0.271	0.4186	0.4198	T KS W BW BT M F
2e-04	0.0024	0.2258	0.0328	0	0.0946	0.8354	BW F W KS BT T M
0.8084	0.0268	0.0014	0.078	0.2228	0.2778	0.56	T W KS BW BT M F
0.8056	0.9808	0.9916	0.8268	0.3276	0.9974	0.9686	BW F KS M W BT T
1	0.9992	0.9998	0.9956	0.9944	1	0.9998	BW KS W (T M) (F BT)
0.8024	0.0248	0.0014	0.075	0.217	0.282	0.565	T W KS BW BT M F
0.8024	0.0248	0.0014	0.075	0.217	0.282	0.565	T W KS BW BT M F

Tabelle 9: Ablehnungsanteile für Simulation 4 je Durchlauf (1-15) für $n_0 = n_1 = 50$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

5.5 Simulation 5 und 6

Die letzten beiden Simulationen unterscheiden sich nur insofern von den Simulationen 3 und 4, als dass sie auch andere kleine Werte abgesehen der Nullen, sprich 1, 2 u.s.w., beinhalten. Man könnte vermuten, dass es in den Teststärken keine gravierenden Unterschiede gibt. Ein Blick auf die Ergebnisse zeigt jedoch ein anderes Bild. Allerdings soll nur auf die größten Unterschiede eingegangen werden.

Die Überlegenheit der Maxsel-Methode ist unübersehbar. Bei den normalverteilten, echt-positiven Werten in Simulation 6 rückt der t-Test als stärkster Test in 5 von 14 Fällen etwas auf, erreicht den Wert von 10 von 14 jedoch nicht annähernd (vgl. Abb. 5). Bei Betrachtung der Ablehnungsanteile in Tabelle 11 und 12 erkennt man an erstgenannter, dass auch hier der BW zusammen mit dem Fisher sehr geringe Ablehnungsanteile hat. Auch eher zu den schwächeren zählt der Kolmogorov-Smirnov-Test. Im Vergleich mit Simulation 4, bei der die kleinen Werte ausschließlich die Nullen beinhalteten, sind alle Tests schwächer geworden, was die Detektierung von Unterschieden betrifft. Sehr gute Ergebnisse erzielen fast alle Tests wieder in den dissonanten Fällen. Vor allem in denen, mit dem größeren Unterschied zwischen den

F	W	T	KS	BW	BT	M	
0	0.0092	0.0156	8e-04	0	0.0058	0.0204	Nullhypothese!
0	0.1352	0.3182	0.0664	0	0.1502	0.3696	(F BW) KS W BT T M
0	0.485	0.9238	0.6634	0	0.7336	0.9512	(F BW) W KS BT T M
0	0.8472	1	1	0	1	1	(F BW) W (T KS BT M)
0.0198	0.0786	0.075	0.0046	0.0018	0.0898	0.0748	BW KS F M T W BT
0.3128	0.2948	0.2668	0.0672	0.0776	0.4442	0.3758	KS BW T W F M BT
0.7856	0.5254	0.504	0.289	0.5018	0.814	0.8722	KS BW T W F BT M
0	0.0016	0.0058	0	0	8e-04	0.0596	(F KS BW) BT W T M
0.128	0.057	0.0122	0.0018	0.0016	0.067	0.0486	BW KS T M W BT F
0	0.0028	0.0564	0	0	0.0064	0.4608	(F KS BW) W BT T M
0.3622	0.0716	0.0062	0.003	0.1074	0.1248	0.251	KS T W BW BT M F
0.5756	0.9238	0.92	0.5646	0.0596	0.9578	0.9182	BW KS F M T W BT
0.9934	0.9916	0.9908	0.9222	0.9352	0.9994	0.997	KS BW T W F M BT
0.7084	1	1	0.9898	0.072	1	1	BW F KS (W T BT M)
1	1	1	0.9994	0.9962	1	1	BW KS (F W T BT M)

Tabelle 10: Ablehnungsanteile für Simulation 4 je Durchlauf (1-15) für $n_0 = n_1 = 25$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

Mittelwerten. Anhand der zweiten Tabelle erkennt man, dass der t-Test gerade in den Durchläufen, in denen konsonantische Situationen vorliegen, und somit die Differenz der Gesamtmittelwerte eher groß ist, diesen Unterschied nicht detektiert. Seine Power liegt hier gerade einmal bei zum Beispiel 0.067 (95% KI, 0.261-0.286). Auch hier scheinen BW und Fisher nicht in der Lage, einen Unterschied in den Verteilungen zu erkennen. Ihre Power liegt in den meisten Fällen unter 0.1.

F	W	T	KS	BW	BT	M	
0.0118	0.0128	0.0166	0.0098	0.002	0.0098	0.0226	Nullhypothese!
0.0118	0.3408	0.6666	0.3466	0.002	0.4454	0.624	BW F W KS BT M T
0.0118	0.8624	0.984	0.9902	0.002	0.9458	0.9996	BW F W BT T KS M
0.0118	0.9758	0.9926	1	0.002	0.9768	1	BW F W BT T (KS M)
0.0808	0.1622	0.1252	0.0796	0.0282	0.129	0.1498	BW KS F T BT M W
0.1668	0.3944	0.3106	0.219	0.0684	0.3366	0.384	BW F KS T BT M W
0.297	0.6554	0.6344	0.4632	0.1396	0.6334	0.7238	BW F KS BT T W M
0.0808	0.007	0.0674	0.0116	0.0282	0.078	0.0908	W KS BW T BT F M
0.1668	0.1146	0.0082	0.087	0.0684	0.098	0.1278	T BW KS BT W M F
0.0808	0.009	0.746	0.044	0.0282	0.5494	0.7582	W BW KS F BT T M
0.1668	0.0364	0.1554	0.0716	0.0684	0.2254	0.3486	W BW KS T F BT M
0.0808	0.8376	0.89	0.7948	0.0282	0.7942	0.941	BW F BT KS W T M
0.1668	0.9524	0.9512	0.9436	0.0684	0.9102	0.9908	BW F BT KS T W M
0.1682	0.0362	0.1554	0.0714	0.0668	0.2344	0.3672	W BW KS T F BT M
0.1682	0.0362	0.1554	0.0714	0.0668	0.2344	0.3672	W BW KS T F BT M

Tabelle 11: Ablehnungsanteile für Simulation 5 je Durchlauf (1-15) für $n_0 = n_1 = 50$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

F	W	T	KS	BW	BT	M	
0.0238	0.0434	0.0468	0.0254	0.0046	0.0482	0.0448	Nullhypothese!
0.073	0.6022	0.6476	0.4146	0.0198	0.5136	0.5494	BW F KS BT M W T
0.093	0.9794	0.9952	0.9702	0.031	0.9774	0.9932	BW F KS BT W M T
0.0942	1	1	1	0.0316	1	1	BW F (W T KS BT M)
0.0454	0.0514	0.0496	0.0356	0.0146	0.0598	0.0774	BW KS F T W BT M
0.069	0.0534	0.0526	0.0472	0.0238	0.0772	0.108	BW KS T W F BT M
0.0994	0.0558	0.053	0.0638	0.038	0.0964	0.1656	BW T W KS BT F M
0.0228	0.2182	0.2738	0.1296	0.0054	0.1754	0.2874	BW F KS BT W T M
0.0392	0.106	0.1456	0.0826	0.0118	0.0946	0.2134	BW F KS BT W T M
0.0226	0.5328	0.7308	0.4546	0.0052	0.5602	0.7922	BW F KS W BT T M
0.0362	0.2248	0.3858	0.1836	0.0108	0.2412	0.4844	BW F KS W BT T M
0.173	0.6398	0.657	0.456	0.0704	0.5812	0.601	BW F KS BT M T W
0.2402	0.6546	0.6584	0.5034	0.1086	0.5984	0.6494	BW F KS BT M W T
0.0344	0.2324	0.3944	0.181	0.0104	0.2454	0.4992	BW F KS W BT T M
0.0344	0.2324	0.3944	0.181	0.0104	0.2454	0.4992	BW F KS W BT T M

Tabelle 12: Ablehnungsanteile für Simulation 6 je Durchlauf (1-15) für $n_0 = n_1 = 50$, die letzte Spalte gibt die Reihenfolge der Tests bzgl. ihrer Stärke an, (...): gleicher Wert.

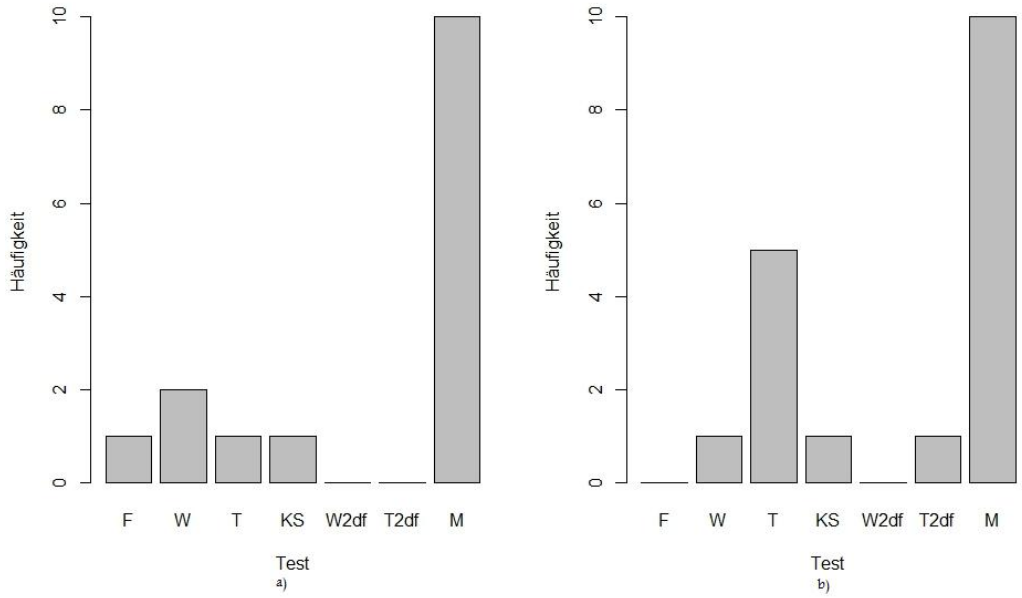

Abbildung 5: Häufigkeiten, wie oft den Tests der maximale Ablehnungsanteil zugeordnet wird für Simulation 5 (a) und Simulation 6 (b) (n=100), falls mehreren Tests der maximale Anteil zugeteilt wird, erhöht sich die Häufigkeit bei jedem entsprechend um 1

6 Fazit

Nach der Einzelbetrachtung jeder Simulation soll in einem kurzen Fazit versucht werden, sich für, aber auch gegen bestimmte Methoden auszusprechen. Da die Tabellen sehr viele Werte enthalten und man leicht den Überblick verlieren kann, zeigt Abb. 6 für jeden Test, wie oft er eine Power von 0.8 überschritten hat. Dieser Wert richtet sich nach Cohen, welcher für den β-Fehler einen viermal so hohen Wert wie für den α-Fehler vorschlägt [Cohen (1988), S.5]. Da das Signifikanzniveau hier jeweils bei $\alpha = 0.05$ lag, ergibt sich ein β-Fehler von 0.2 und damit eine Power, $1-\beta$, von 0.8. Ergänzend ist dieser Wert anteilig an den Simulationsdurchläufen dargestellt, da der Fisher, BW und BT in Simulation 2 nicht vertreten waren. Die Maxsel-Methode erreichte 50 mal eine Power > 0.8, was einem Anteil von 33% entspricht. Der BT überschreitet diesen Wert 44 mal, dies bedeutet in 32% der Fälle.

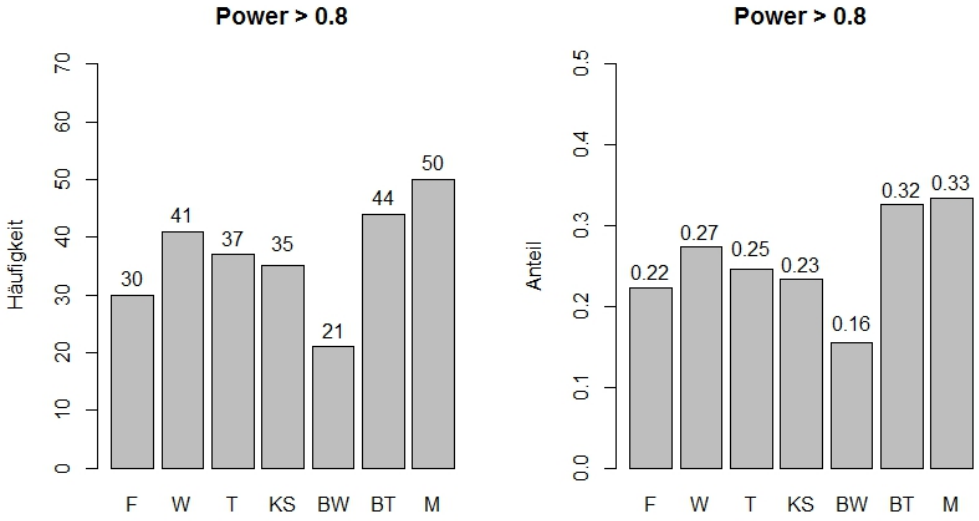

Abbildung 6: Absolute (rechts) und relative (links) Häufigkeiten für Power > 0.8.

Über sein Verhalten in Simulation 2 ist jedoch nichts bekannt, daher ist die Maxsel-Methode vorzuziehen.

Basierend auf den Ergebnissen, scheint es, als sei die Maxsel-Methode gut für bestimmte Anwendungen geeignet, da sie konstant sehr stark ist. In den wenigen Fällen, in denen ein anderer Test besser ist, ist der Unterschied nur gering und hat die Maxsel-Methode eine geringe Power, so sind die übrigen Tests auch nicht (viel) besser. Somit ist man mit Maxsel am ehesten auf der sicheren Seite, was die Power betrifft. Von BW, Fisher und auch BT Test ist eher abzuraten. Die beiden zuerst genannten sind die schwächsten aller miteinander verglichenen Tests. Beim BT Test ergibt sich dasselbe Problem wie auch beim Fisher und BW: Es muss ein (bzw. mehrere) Schwellenwert(e) vor Durchführung des Tests bestimmt werden. Ob diese jedoch geeignet sind, ist zuvor meist schwierig zu bestimmen. So ist auch bei einer Power von 1 beim Fisher-Test nicht genau zu sagen, ob der erkannte Unterschied in den Gruppen nur aufgrund des Cutpoints gemacht wurde und ob das Ergebnis bei einem anderen Cutpoint nicht komplett anders wäre.

Wilcoxon-, t- und Kolmogorov-Smirnov-Test liegen nicht weit auseinander. Da der Wilcoxon-Test ein Test ist, der damit arbeitet, dass Verteilungen sich nur bzgl. der Lage unterscheiden [vgl. Toutenburg und Heumann (2008), S. 174], erkennt er die Unterschiede, die aufgrund eines Unterschieds in den Nullanteilen vorliegen, nur schwer. Liegt jedoch (zusätzlich) auch eine Verschiebung bedingt durch die Mittelwerte vor, so ist er relativ stark. Dasselbe gilt für den t-Test. Der Kolmogorov-Smirnov-Test detektiert Unterschiede nur, falls sie groß sind, unabhängig von der Ursache.

Die Maxsel-Methode, die sich für jede Situation ihren Cutpoint anhand der maximalen χ^2-Statistik neu berechnet (unbeeinflusst durch *fishing for significance* [1]), scheint somit, wie bereits erwähnt, am besten geeignet, um Daten, die hohe Anteile an Nullen oder auch kleinen Werten beinhalten, auf Unterschiede und somit eine Abhängigkeit, zu untersuchen. In den Simulationen erwiesen sich Stichprobengrößen > 20 als geeignet dafür, dass die Tests eine höhere Power erreichten. Es ist daher auch in der Praxis in Erwägung zu ziehen, die Gruppen lieber etwas größer zu wählen um unnötig hohe β-Fehler, und somit eine geringe Power, zu vermeiden.

[1] es wird der Cutpoint ausgewählt, mit welchem der kleinste p-Wert resultiert

7 Anwendung in der Praxis

Für eine praktische Anwendung der Tests wurde ein Datensatz über Krebspatienten betrachtet. Es wurden der t-Test, der Wilcoxon-Test, der Kolmogorov-Smirnov-Test und die Maxsel-Methode angewandt und die Ergebnisse kritisch verglichen. Fisher, BW und BT wurden nicht gerechnet, da eine Dichotomisierung der Variablen beim Wert 0 nicht geeignet ist. Grund dafür ist, dass einige Variablen keine Ausprägung "0" besitzen und die Tests somit nicht durchführbar sind.

7.1 Überblick über die Daten

Der Datensatz enthält Informationen über Krebspatienten. Die für die Fragestellung dieser Arbeit interessierenden Variablen sind *Metastasiert* (Waren Metastasen vorhanden?), *TodTUassoziiert* (War der Tod des Patienten tumorbedingt?) und die Variablen, welche die Methylierungs-Prozentwerte der einzelnen Cytosin-Guanosin-Basenfolgen (CpGs) der Gene angeben. *Metastasiert* und *TodTUassoziiert* sind ursprünglich binär kodiert mit den Faktoren *Ja* und *Nein*. Um die Tests anwenden zu können, wurde *Ja* in 1 und *Nein* in 0 geändert. Die Häufigkeiten in den Gruppen sind in folgender Tabelle dargestellt:

	Tumor = 0	Tumor = 1	
Meta = 0	26	0	26
Meta = 1	6	16	22
Summe	32	16	48

Für jedes Gen gibt es eine Spalte, die den Methylierungswert bei gesunden Menschen angibt. Eine weitere Spalte gibt zusätzlich den Mittelwert der einzelnen Methylierungs-Prozentwerte für jeden Patienten pro Gen an. Diese sind bei der Testberechnung jedoch nicht von großem Interesse, da sie nicht die kleinen Werte nahe der Null, welche in dieser Arbeit interessieren, enthalten. Das Beispiel in Abb. 7, welches ausgewählt wurde, da es für die vorliegende Fragestellung gut passt, da viele kleine Werte enthalten sind, soll

veranschaulichen, wieso diese Art von Daten hier besonders interessant ist. Man erkennt sogleich die stark linkssteile Verteilung der Datenwerte, welche bei den verbleibenden Genen einen ähnlichen Charakter aufweisen.

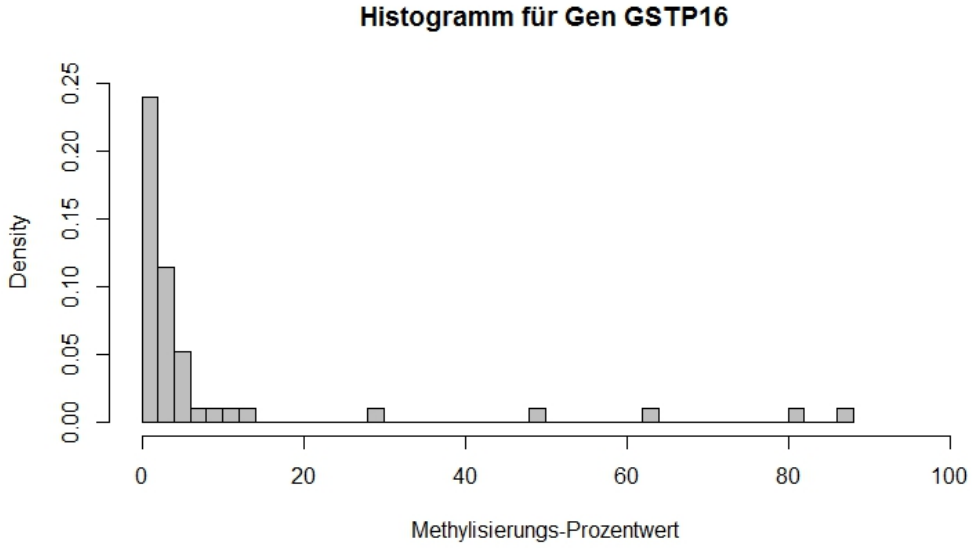

Abbildung 7: Histogramm der Methylierungs-Prozentwerte des CpGs GSTP16.

7.2 Ergebnisse

Der Wilcoxon-Test, der t-Test, der Kolmogorov-Smirnov-Test und die Maxsel-Methode wurden angewandt, um zu überprüfen, ob ein Zusammenhang zwischen einer stattgefundenen Metastasierung bzw. einem durch den Tumor verursachten Tod und den Methylierungs-Prozentwerten an den verschiedenen CpGs der einzelnen Gene vorhanden ist. Das Signifikanzniveau wurde mit $\alpha = 0.05$ festgelegt. Als erstes wurden die Tests bezüglich der binären Variable *Metastasiert* durchgeführt. Abb. 8 zeigt die Häufigkeit eines p-Wertes < 0.05 je Test. Es fällt sofort der t-Test auf, welcher nur bei einem einzigen CpG einen Unterschied in den Verteilungen der Methylierungs-Prozentwerte für Patienten mit und ohne Metastasen erkennt. Doch wie oft gibt es eine Übereinstimmung zwischen den Tests bezüglich der CpGs, für welche sie

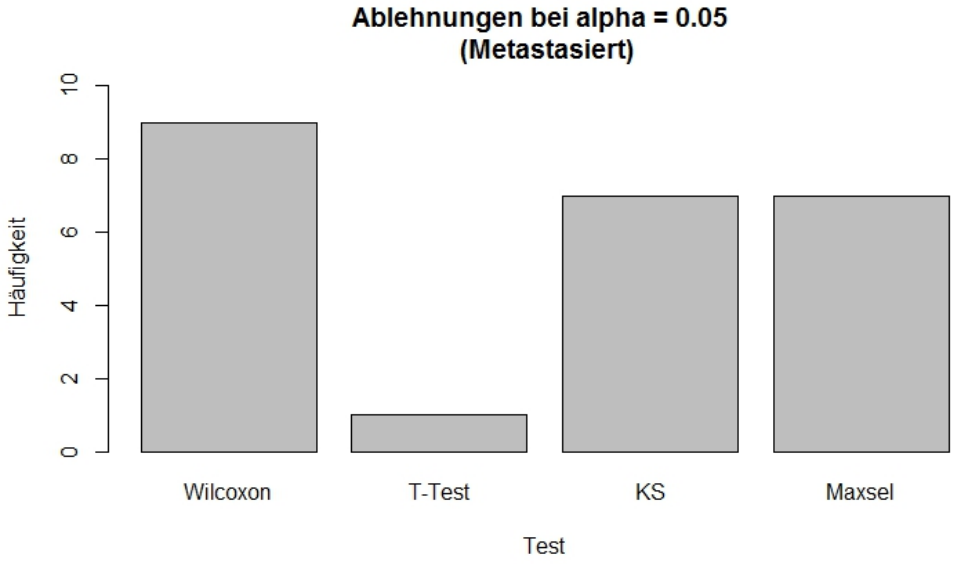

Abbildung 8: Anzahl Ablehnungen der Nullhypothese ("Die Verteilungen in den Gruppen sind identisch") je Test für "Metastasiert".

einen Zusammenhang ausschließen? Eine Antwort gibt Tabelle 13.

Das einzige CpG, bei dem der t-Test einen Zusammenhang verwirft, ist *APC.3*. Der p-Wert ist 0.04. Der Wilcoxon ist mit einem nahezu identischen p-Wert (Werte sind gerundet) der einzige weitere Test, der auch die Nullhypothese ablehnt. Zieht man Parallelen zu den Simulationsergebnissen der vorherigen Kapitel, so findet man diese zum Beispiel bei Simulation 1, Durchlauf 9 (vgl. Tabelle 4). Auch die Gruppengrößen stimmen in etwa überein. In diesem Fall läge hier sowohl ein Unterschied in den Nullanteilen (bzw. bei den kleinen Werten) als auch in den Mittelwerten der Werte, die größer Null sind, vor. Bleibt man beim t-Test, so fällt auf, dass die übrigen p-Werte alle sehr hoch sind. Das bedeutet, dass er den Zusammenhang nur zu sehr hohen (nicht mehr sinnvollen) Signifikanzniveaus ablehnt. Auch, und gerade, bei den Variablen, für welche die anderen Tests die Nullhypothese sicher ablehnen. Ein weiterer, auffälliger Punkt ist, dass die Maxsel-Methode den Unterschied bei der Hälfte aller für ohne Zusammenhang befundenen Variablen detektiert. In den Fällen, in denen sie die Nullhypothese nicht ablehnt, ist der p-Wert nahe der 0.05. Bei diesen Variablen entscheidet ausnahmslos

	wilcox	ttest	ks	maxsel
APC_mean	0.02	0.12	0.03	0.04
APC	0.14	0.23	0.05	0.04
APC.3	0.04	0.04	0.14	0.06
APC.5	0.01	0.13	0.07	0.06
APC.6	0.02	0.82	0.05*	0.05
DAPK.12	0.04	0.10	0.15	0.11
GADD45a.2	0.03	0.62	0.05	0.05
GADD45a.3	0.01	0.30	0.04	0.03
GSTP12	0.05*	0.39	0.15	0.18
p14.16	0.08	0.97	0.04	0.01
p73.1	0.12	0.53	0.02	0.02
Endoglin10	0.35	0.60	0.13	0.05*
Endoglin13	0.21	0.44	0.04	0.03
Endoglin16	0.03	0.11	0.07	0.05

Tabelle 13: p-Werte für die CpGs, bei denen mindestens ein Test H_0 ablehnt (Metastasiert), * Werte auf 0.05 aufgerundet (korrekter Wert < 0.05)!.

der Wilcoxon gegen H_0.

Das Gleiche gilt für die Variable *Tumorassoziiert*. Die Ablehnungshäufigkeiten sind in Abb. 9 zu finden. Bei den Variablen *APC_mean, APC.5, GSTP2_mean* und *RAS.1* ist der t-Test der einzige Test, der die Nullhypothese nicht ablehnt. Zieht man Parallelen zu den Simulationsergebnissen der vorherigen Kapitel, so ist dies mit Durchlauf 7 von Simulation 3 (für $n = 50$) bzw. mit den Durchläufen 9 und 11, ebenso von Simulation 3, zu vergleichen (s. Tabelle 7 und 8). Da in dem aktuellen Datensatz Gruppengrößen von $n \approx 50$ vorliegen, ist erstgenannte Situation besser zum Vergleich geeignet. So besteht die Möglichkeit, dass der Unterschied in den Verteilungen zwar sehr groß ist, dies jedoch nur aufgrund der großen Differenzen in den Nullanteilen zustande kommt. Da der t-Test die Klumpung bei der Null (zumindest hier) ignoriert, entscheidet er für Gleichheit. Tatsächlich liegt der Mittelwert für z.B. *APC.5* in der Gruppe der Patienten mit Metastasen (*Metastasiert=1*) bei 11.872, in der Gruppe der Patienten ohne Metastasen (*Metastasiert=0*) bei gerade einmal 4.42.

APC.3 ist die einzige Variable, für die der t-Test die Hypothese des Zu-

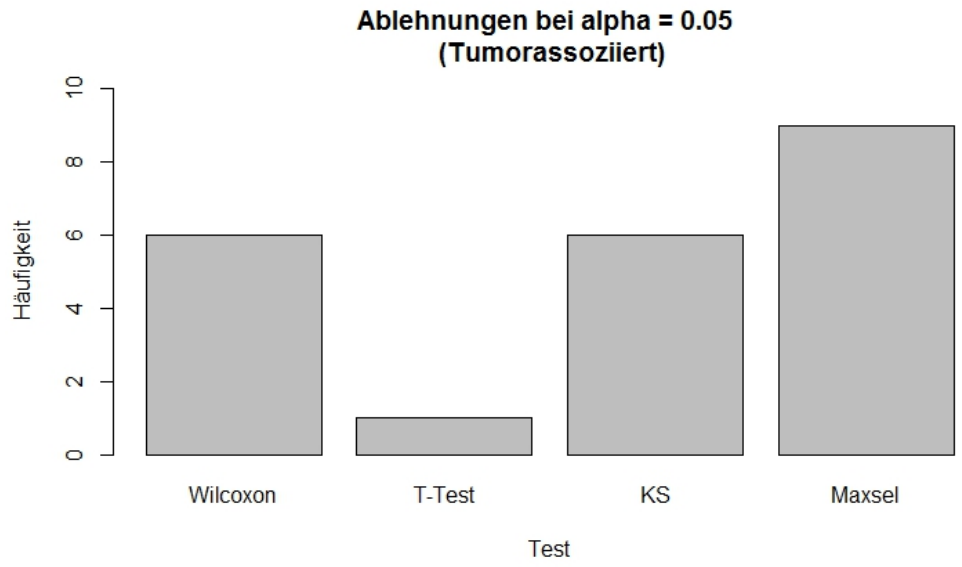

Abbildung 9: Anzahl Ablehnungen je Test für "Tumorassoziiert".

sammenhangs verwirft und somit einen Unterschied in den Verteilungen detektiert. Die Maxsel-Methode (zuvor als äußerst zuverlässig befunden) bekräftigt diese Entscheidung. Eine Übereinstimmung könnte hier mit Durchlauf 3 von Simulation 4 (Tabelle 10) gesehen werden. Hier hatten sowohl t-Test als auch Maxsel einen hohen Ablehnungsanteil und der Kolmogorov-Smirnov-Test einen höheren Ablehnungsanteil als der Wilcoxon. Dies spricht dafür, dass der p-Wert des Kolmogorov-Smirnov-Tests (0.06) kleiner als der des Wilcoxon (0.07) ist.

	wilcox	ttest	ks	maxsel
APC_mean	0.02	0.08	0.03	0.02
APC	0.11	0.14	0.02	0.01
APC.3	0.07	0.04	0.06	0.01
APC.5	0.02	0.10	0.03	0.01
APC.7	0.09	0.10	0.10	0.02
GSTP2_mean	0.04	0.65	0.03	0.03
GSTP15	0.03	0.52	0.06	0.08
p14.5	0.04	0.21	0.10	0.17
RAS.1	0.02	0.46	0.03	0.02
RAS.2	0.05	0.45	0.03	0.02
RAS.4	0.59	0.29	0.37	0.03

Tabelle 14: p-Werte für die Gene, bei denen mindestens ein Test einen Zusammenhang ausschließt (Tumorassoziiert),* Werte auf 0.05 aufgerundet!.

8 Zusammenfassung und Ausblick

In dieser Arbeit wurde die Maxsel-Methode mit weiteren statistischen Tests verglichen. Der Vergleich geschah bezüglich ihrer Fähigkeit, Zusammenhänge zwischen einer binären Variable und einer Variable, welche viele kleine Werte und vor allem Nullen enthält, zu erkennen. Diese Tests waren der Wilcoxon-Test, der t-Test, der Kolmogorov-Smirnov-Test und die sogenannten Two-Part-Models bestehend aus Binomial- und Wilcoxon- bzw. t-Test. Dazu wurden zunächst mehrere Simulationen durchgeführt. Da hier Vorwissen über die Situation in den Daten vorlag, konnte anhand der Ablehnungsanteile die Power der Tests in den jeweiligen Fällen bestimmt werden. Diese unterschied sich sowohl unter den Tests, als auch für die jeweiligen Durchläufe. Getestet wurde für verschiedene Verteilungen derjenigen Werte der Gruppe, welche größer Null waren (bzw. nicht zu den kleinen Werten zählten), verschiedene Nullanteile und verschiedene Mittelwerte. Es wurde festgestellt, dass die Maxsel-Methode in 33% der Fälle die Nullhypothese mit einer Power > 0.8 abgelehnt hat. Danach folgte der BT mit 32% und der Wilcoxon mit 27%. Dieses Ergebnis deckt sich mit dem der Tests an dem Datensatz über Krebspatienten. Hier war von Interesse, ob die Tests jeweils einen Zusammenhang zwischen den binären Variablen *Metastasiert* und *Tumorassoziiert*

und den Methylierungs-Prozentwerten verschiedener Gene detektieren oder ablehnen. Es erzielte jeweils einmal der Wilcoxon und einmal die Maxsel-Methode die meisten Ablehnungen der Nullhypothese (kein Unterschied in den Verteilungen der Gruppen). Diese ergänzten sich insofern gut, als dass, falls die Maxsel-Methode einen Zusammenhang nicht ausschließen konnte, der Wilcoxon-Test dies jedoch tat. Folglich scheint die Maxsel-Methode in Verbindung mit dem Wilcoxon-Test ein geeignetes Mittel zu sein, um Daten, welche viele kleine Werte enthalten, und somit stark linkssteil verteilt sind, auf Zusammenhänge zu untersuchen. Da die Situation in den reellen Daten nicht im Voraus bekannt ist, und die Tests unterschiedlich gut darauf reagieren, scheint es eher nicht geeignet, sich auf nur eine Methode zu verlassen. Der BT ist, wie auch der BW und Fisher, trotz einer Power > 0.8 in 32% der Fälle, nicht immer geeignet, da hier im Voraus ein Cutpoint für die Dichotomisierung der nicht binären Variable bestimmt werden muss.

Interessant könnte nun noch sein, die Simulationen für unterschiedliche Gruppengrößen (z.B. $n_0 = 30, n_1 = 70$) durchzuführen und weitere Kombinationen in Nullanteilen und Mittelwerten zu untersuchen. Da die Gruppengrößen in der Praxis häufig nicht gleich groß sind, wäre es interessant zu wissen, wie sich die verschiedenen Tests in solchen Situationen verhalten und ob es auch Tendenzen zu einem bestimmten Verhaltensmuster gibt, anhand welchem man sich entscheiden kann, welcher Test in entsprechenden Situationen eher verwendet werden sollte. Denn möglicherweise verändert sich das Verhalten der Tests durch diese Veränderung. Weiter könnte zusätzlich ein Two-Part-Modell bestehend aus Binomial- und Kolmogorov-Smirnov-Test mit aufgenommen werden. Es ist jedoch zu erwarten, dass dessen Ergebnisse bzgl. der Power ähnlich denen von BT und BW sind, da sich der Kolmogorov-Smirnov-Test nicht sonderlich stark von Wilcoxon- und t-Test unterschieden hat. Auch das Problem der Cutpoint-Findung läge hier vor. Da sich die Ergebnisse verändert haben, falls die echt-positiven Werte anders verteilt waren, könnten hier noch Simulationen mit weiteren Verteilungen, wie zum Beispiel der log-Gammaverteilung, durchgeführt werden. Auch könnten die Gruppengrößen für die Simulationen 5 und 6 verkleinert und die Ergebnisse mit denen der größeren Gruppen verglichen werden.

Literatur

Boulesteix AL (2006). "Maximally Selected Chi-square Statistics for Ordinal Variables." *Biometrical Journal*, **48**, 451–462. doi:10.1002/bimj.200510161. URL `http://dx.doi.org/10.1002/bimj.200510161`.

Cohen J (1988). *Statistical Power Analysis for the Behavioral Sciences - Second Edition*. Lawrence Erlbaum, New Jersey.

Fahrmeir L, Künstler R, Pigeot I, Tutz G (2010). *Statistik: Der Weg zur Datenanalyse*. Springer Verlag, Berlin Heidelberg.

Lachenbruch PA (2001). "Comparison of two-part models with competitors." *Statistics in Medicine*, **20**, 1215–1234. doi:10.1002/sim.790. URL `http://dx.doi.org/10.1002/sim.790`.

Lachenbruch PA (2002). "Analysis of data with excess zeros." *Statistical Methods in Medical Research*, **11**, 297–302. doi:10.1191/0962280202sm289ra. URL `http://dx.doi.org/10.1191/0962280202sm289ra`.

Leonhart R (2009). *Lehrbuch Statistik - Einstieg und Vertiefung*. Huber Verlag, Bern.

Steland A (2010). *Basiswissen Statistik*. Springer Verlag, Berlin Heidelberg.

Toutenburg H, Heumann C (2008). *Induktive Statistik - Eine Einführung mit R und SPSS*. Springer Verlag, Berlin Heidelberg.

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 23. Juli 2013	Eva-Maria Müntefering