Ludwig-Maximilians-Universität
München

Institut für Statistik

Simulationsstudie zum Gütevergleich ausgewählter Hypothesentests unter potentiell problematischen Datensituationen

Betrachtung von Wilcoxon-Vorzeichen-Rang-, Vorzeichen- und t-Test im Einstichprobenfall

---

Bachelorarbeit

Zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)

Autor:
Tobias Steinherr

Matrikelnummer:
****

Betreuer:
Paul Fink, M.Sc.

11. März 2015

Zusammenfassung

In der vorliegenden Arbeit werden drei verschiedene Hypothesentests anhand unterschiedlicher Daten bezüglich ihrer Güte verglichen. Die untersuchte Testproblematik bezieht sich auf den einseitigen Einstichprobenfall. Als Tests wurden der Wilcoxon-Vorzeichen-Rang-Test, der Vorzeichentest und der t-Test ausgewählt, wobei die beiden zuerst genannten nichtparametrische Verfahren darstellen.

Zunächst wird die Testproblematik näher geschildert sowie die Methodik der jeweiligen Verfahren näher geklärt. Außerdem wird die Güte erklärt und geschildert, wieso zum Vergleich randomisierte Tests herangezogen werden.

Im Anschluss darauf wird dazu das Prinzip von Simulationen, mit Hilfe derer die Gütefunktionen der Tests bestimmt werden sollen, geschildert, woraufhin diese durchgeführt werden. Neben normalverteilten Daten werden daraufhin Daten unterschiedlichster theoretischer Verteilungen betrachtet, die vor allem die an den t-Test gebundenen Voraussetzungen nicht erfüllen. Zu allen Datensituationen werden Gütefunktionen grafisch dargestellt.

Zudem wird mit der Simulation der finiten relativen Effizienz versucht, eine Kennzahl für die Güteunterschiede der Tests untereinander und diejenigen innerhalb der verschiedenen Verteilungen zu erhalten, die einen Vergleich einfacher und übersichtlicher machen soll.

Inhaltsverzeichnis

1 Einleitung

Um eine bestimmte statistische Problematik zu untersuchen, existieren dazu meist zahlreiche verschiedene Möglichkeiten, sodass es Schwierigkeiten bereiten kann, sich für eine zu entscheiden. Auch in der Fragestellung, ob sich im Einstichprobenfall ein Mittelwert in der Grundgesamtheit von einem konkreten hypothetischen Wert unterscheidet, wird man vor eine beachtliche Auswahl an Tests gestellt.

Der vermutlich bekannteste und am häufigsten verwendete Test hierfür ist der t-Test im Einstichprobenfall, der jedoch an gewisse Voraussetzungen gebunden und in seiner Methodik weniger simpel ist. Ohne größere Voraussetzungen kommen der Vorzeichen- und der Wilcoxon-Vorzeichen-Rang-Test aus, die in ihrer Handhabung zudem intuitiver und einleuchtender erscheinen.

Die Frage, die sich nun stellt, ist, wie effizient diese verschiedenen Methoden im Vergleich zueinander sind. Wie sehr kann beispielsweise ein Vorzeichentest, der in seiner Durchführung ohne Weiteres einem Fachfremden mit etwas Sinn für Mathematik und Stochastik begreiflich näher gebracht werden kann, mit einem t-Test mithalten, der, gerade was die Verteilung seiner Teststatistik betrifft, durchaus komplizierter erscheint? Oder fällt der t-Test bei Missachtung seiner Voraussetzungen an die Daten der Stichprobe wirklich zwingend ab und wenn ja, wie sehr? In welcher Situation ist welcher Test der geeignetste?

Zu solchen Fragen eine allgemeingültige Antwort zu liefern, wird aller Voraussicht nach nie ganz möglich sein. Diese Arbeit ist eine Studie, die anhand von Simulationen und unterschiedlichsten Datensituationen konkrete Einblicke in die Güte der drei betrachteten Tests geben und eventuelle Problematiken der einzelnen Tests herausarbeiten und veranschaulichen wird.

2 Theorie

2.1 Testproblematik

Hypothesentests im Einstichprobenfall, wie sie hier besprochen werden, überprüfen grundsätzlich die Fragestellung, ob sich der Mittelwert einer Variable X in der Grundgesamtheit von einem hypothetischen Wert μ_0 unterscheidet. Anhand von Stichproben werden verschiedene Tests die folgenden Hypothesen gegeneinander abwägen:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Eine eindeutige Entscheidung für eine der beiden Hypothesen zu fällen, ist meist nicht möglich, da dafür für jedes einzelne Element der Grundgesamtheit die interessierende Variable zu messen wäre. Hier kommen Hypothesentests ins Spiel. Es werden für eine Stichprobe zufällig eine bestimmte Anzahl n Elemente aus der Grundgesamtheit herausgegriffen und deren Variablenwerte erfasst. Das weitere Vorgehen anhand der gewonnenen Daten ist von Test zu Test unterschiedlich und wird in Abschnitt 2.2 zu allen Tests vorgestellt. Grundsätzlich werden die Werte der Stichprobe mit dem hypothetischen Wert μ verglichen, in welcher Form auch immer. Je nachdem, wie das Ergebnis aussieht, kann zumindest die Sicherheit abgewogen werden, ob μ in der Grundgesamtheit von μ_0 abweicht oder nicht (Fahrmeir et al.; 2007, S. 397).

In dieser Arbeit werden ausschließlich symmetrische Verteilungen beobachtet und deshalb wurde entschieden, die Hypothesen auf eine Seite zu beschränken. Es wird getestet, ob der Mittelwert einer Stichprobe kleiner als ein hypothetischer Wert ist oder nicht. Das heißt, dass die Hypothesen, die überprüft werden, genau genommen folgender Gestalt sind:

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

Für μ_0 wird dabei in der vorliegenden Arbeit immer der Wert 0 gewählt, doch dazu später mehr.

2.2 Die Tests

In den folgenden Abschnitten sollen die Methodiken der untersuchten Tests dargestellt werden. Vor allem bei den nonparametrischen Tests soll die Erklärung zur Bildung der Teststatistik und die damit verbundene Testentscheidung ausführlicher sein. Ebenso sollen unterschiedliche Vorgehensweisen zur Behandlung von Problematiken aufgeführt werden.

2.2.1 Vorzeichentest

Der Vorzeichentest zählt zu den nichtparametrischen Verfahren und ist in seiner Methodik sehr einfach nachzuvollziehen. Für seine Durchführung sind an die Daten lediglich die Voraussetzungen Stetigkeit, Symmetrie und Unabhängigkeit gebunden, wobei auch die Stetigkeit in der Praxis nicht von größerer Bedeutung ist (siehe Duller (2008, S. 135)). Gegeben sei eine Stichprobe X der Länge n, also $X = \{x_1, x_2, x_3, ...x_n\}$. Es soll wie erwähnt überprüft werden, ob die Ausprägungen der Stichprobe im Mittel signifikant kleiner als ein hypothetischer Wert μ_0 sind. Dazu werden zunächst sämtliche Differenzen $D_i = x_i - \mu_0$ gebildet. Die Teststatistik entspricht nun der Anzahl derjenigen D_i, die die Bedingung $sign(D_i) = 1$ erfüllen, also der Anzahl an Ausprägungen von X, die größer als der hypothetische Wert μ_0 sind (Büning & Trenkler; 1994, S. 93). Diese Teststatistik wird im Folgenden A genannt und damit wird nun überprüft, ob der Median kleiner als μ_0 ist oder nicht bzw. für welche der beiden Hypothesen sich entschieden wird:

$$H_0 : x_{med} \geq \mu_0$$
$$H_1 : x_{med} < \mu_0$$

Wie kann das mit der Teststatistik A überprüft werden? Da für diese nur von Interesse ist, ob die Werte von X größer oder kleiner als μ_0 sind, ist es gleichbedeutend mit der Alternativhypothese, dass die Wahrscheinlichkeit für ein positives D_i weniger als 50% beträgt. Sei k nun eine bestimmte Ausprägung der Teststatistik A, so berechnet man die Wahrscheinlichkeit für bis zu k positive Vorzeichen, wenn unterstellt wird, dass es rein zufällig ist, ob D_i nun positiv ist oder nicht. Dies ist zu berechnen mit der Verteilungsfunktion der Binomialverteilung $B(k, n, 0.5)$ und zwar folgendermaßen (Fahrmeir et al.; 2007, S. 254):

$$\mathbb{P}(A \leq k) = \sum_{i=0}^{k} \binom{n}{i} 0.5^i (1 - 0.5)^{n-i} = (\sum_{i=0}^{k} \binom{n}{i}) \cdot 0.5^n$$

Für die Extremwerte $k = n$ (jedes Element von X ist größer als μ_0) oder $k = 0$ (kein Element von X größer als μ_0) ergeben sich die Werte 1 und 0.5^n. Das bedeutet, dass der niedrigste zu erreichende p-Wert eines einseitigen Vorzeichentests gleich 0.5^n ist und die Nullhypothese auf einem Niveau von $\alpha < 0.5^n$ nie verworfen werden kann.

Nun gibt es die Möglichkeit von Nulldifferenzen, was heißt, dass ein Wert der Stichprobe genau dem von μ_0 entspricht. Für diesen Fall werden zwei Möglichkeiten, damit umzugehen, genannt. Die erste Option ist es, diese Werte aus der Stichprobe zu entfernen und nur die Werte zu betrachten, die sich von μ_0 unterscheiden (Büning & Trenkler; 1994, S. 94). Infolgedessen verkleinert sich der Stichprobenumfang entsprechend und eine Möglichkeit wird dabei eventuell außer Acht gelassen: Angenommen eine Stichprobe vom Umfang $n = 50$ weist folgende Werte auf: Zwei Werte sind größer als das hypothetische μ_0, zehn sind kleiner und die restlichen 38 sind genau so groß wie μ_0. 38 Werte fallen aus

der Stichprobe und für $P(A \leq k)$ ergibt sich $(\sum_{i=0}^{2} \binom{12}{i}) \cdot 0.5^{12} \approx 0.0193$, was bedeutet, dass das Testergebnis auf dem 5%-Signifikanzniveau aussagt, dass der Median kleiner als μ_0 ist. Nun entspricht aber ein Großteil der Werte exakt μ_0, was die Vermutung nahelegt, dass sich der Median in Wahrheit eben nicht von μ_0 entscheidet.
Abhilfe kann hierbei eine weitere Methode im Umgang mit Nulldifferenzen schaffen. Hier fallen die Werte mit Nulldifferenzen nicht aus der Stichprobe, sondern es wird per Zufall entschieden, ob jedem der entsprechenden Werte entweder ein positives oder negatives Vorzeichen zugeordnet wird (Büning & Trenkler; 1994, S. 94 & 95). In diesem Beispiel wären dadurch zusätzliche 19 negative und 19 positive Vorzeichen im Mittel zu erwarten, sodass sich in diesem Fall ein p-Wert von $(\sum_{i=0}^{21} \binom{50}{i}) \cdot 0.5^{50} \approx 0.1611$ ergeben würde. Dabei würde die Nullhypothese bei den gleichen ursprünglichen Daten wie vorhin weder auf dem 10%-, geschweige denn dem 5%-Signifikanzniveau abgelehnt werden. Zu dieser Methode ist zu sagen, dass sie die Daten mehr oder minder verfälscht. Andererseits ist zu erwähnen, dass die Sinnhaftigkeit dahinter auch darin liegt, dass eigentlich von stetigen Daten ausgegangen wird, was bedeutet, dass die Wahrscheinlichkeit für zwei oder mehrere gleiche Werte theoretisch gleich 0 ist. In der Praxis kann es jedoch beispielsweise zu Messungenauigkeiten kommen, sodass diese Wahrscheinlichkeit dann eben doch gegeben ist. Wenn also vorausgesetzt wird, dass zwei oder mehrere Werte nicht gleich gleich groß sein können, genau dies aber auftritt, so kann oder muss sogar davon ausgegangen werden, dass ein Fehler dahintersteckt und so erscheint die Vergabe von zufälligen Vorzeichen als durchaus legitim.

2.2.2 Wilcoxon-Vorzeichen-Rang-Test

Nicht nur vom Namen, sondern auch von seinem Vorgehen her ist der Wilcoxon-Vorzeichen-Rang-Test nicht ganz unterschiedlich zum eben vorgestellten Vorzeichentest. Neben den gleichen Voraussetzungen wie für den Vorzeichentest (siehe Abschnitt 2.2.1 und Duller (2008, S. 135)) benötigt er zum Berechnen der Teststatistik alle Differenzen zwischen den Werten von X und μ_0, also $D_i = x_i - \mu_0$. Diese Differenzen werden nun betragsmäßig der Reihe nach geordnet und die ursprünglichen Werte durch die Ränge 1 (Wert mit der geringsten Distanz zu μ_0) bis n (größte Differenz) ersetzt. Zusätzlich dazu wird jeder Rang $rg(x_i)$ mit dem gleichen Vorzeichen wie dem des entsprechenden D_i versehen; aus einer negativen Differenz folgt ein negativer Rang (Fahrmeir et al.; 2007, S. 443). Für die Teststatistik W^+ werden nun alle positiven Ränge addiert, also

$$W^+ = \sum_{i=1}^{n} rg(|D_i|)Z_i$$

mit

$$Z_i = \begin{cases} 1, & D_i > 0 \\ 0, & \text{sonst} \end{cases}$$

Zu dieser Teststatistik kann gesagt werden, dass ihr Minimum bei 0 liegt (alle Differenzen negativ) und ihr Maximum bei $\sum_{i=1}^{n} = \frac{n(n+1)}{2}$ (alle Differenzen

positiv). Für den einseitigen Test mit der Alternativhypothese $H_1 : x_{med} < \mu_0$ ist nun von Interesse, wie hoch die Wahrscheinlichkeit für eine Teststatistik ist, die maximal so groß ist wie die sich ergebende. Sei w das Ergebnis eines Tests, so muss also die Wahrscheinlichkeit $P(W^+ \leq w)$ berechnet werden. Da dies neben einigem Rechenaufwand gerade bei größerem Stichprobenumfang auch erheblich viel Kombinatorik erfordert und die Methodik im Rahmen dieser Arbeit von untergeordnetem Interesse ist, sei an dieser Stelle auf Wilcoxons Paper verwiesen (Wilcoxon; 1945). Grundsätzlich ist für die Antwort darauf, wie viele Möglichkeiten es für das Zustandekommen einer bestimmten Rangsumme gibt, immer folgendes von Interesse: Wie viele Möglichkeiten gibt es, diese Rangsumme als Summe aus bis zu n unterschiedlichen, positiven und natürlichen Zahlen kleiner oder gleich n darzustellen? Für den Fall, dass die Rangsumme gleich 0 ist, gibt es nur eine Möglichkeit, und zwar, dass kein Rang positiv ausfällt (Wilcoxon; 1945).

Um sich diese enorme Arbeit an Kombinatorik zu ersparen, ist die Teststatistik W^+ auch in einer Vielzahl an Tabellen festgehalten (siehe z.B. Fahrmeir et al. (2007, S. 590)). So kann auf einem Niveau von α von der Alternativhypothese ausgegangen werden, wenn die Teststatistik kleiner als das α-Quantil der tabellierten Verteilung von W^+ ausfällt.

Im Fall des Wilcoxon-Vorzeichen-Rang-Tests kann es nun ebenso zu Nulldifferenzen kommen. Auch hier ist es eine Möglichkeit, diese aus der Stichprobe zu eliminieren (Büning & Trenkler; 1994, S. 98). Wenn sich dazu nicht entschlossen wird, würde dem betroffenen Wert der niedrigste Rang zugeordnet werden, da die Differenz von 0 die kleinstmögliche Absolutdifferenz darstellt. Da der Wert 0 jedoch weder positiv noch negativ ist, stellt sich entsprechend die Frage, ob seinem Rang überhaupt ein bestimmtes Vorzeichen zugeordnet werden kann. Eine Möglichkeit, die in dieser Arbeit betrachtet wird, ist diejenige, dass der entsprechende Rang halbiert und zur einen Hälfte als positiv und zur anderen als negativ angesehen wird, was für die Teststatistik bedeutet, dass die Summe der positiven Ränge um einen halben Rang addiert wird. Bei mehreren Nulldifferenzen sei d deren Anzahl; dann wird jedem der dazugehörigen Werte der Rang $\frac{d+1}{2}$ zugeordnet (Bortz et al.; 2008, S. 262).

Eine weitere Schwierigkeit, die sich beim Wilcoxon-Vorzeichen-Rang-Test zusätzlich ergeben kann, ist die Möglichkeit von Rangbindungen. Diese sind gegeben, wenn zwei oder mehrere Absolutdifferenzen den gleichen Betrag aufweisen, also wenn zwei Werte der Stichprobe den gleichen Abstand zum hypothetischen Wert μ_0 haben. Diese Bindungen (englisch: Ties) können auf unterschiedlichste Arten behandelt werden. Beispielsweise können zufällige Ränge vergeben werden. Die geläufigste Methode ist es, dass der Durchschnitt der eigentlichen Ränge vergeben wird; die Methode, die in dieser Arbeit betrachtet wird. Dies entspricht bei Überlegung genau der vorgestellten Methode für diejenige Behandlung von Nulldifferenzen, in der jeder der d Nulldifferenzen der Rang $\frac{d+1}{2}$ zugewiesen wird (Bortz et al.; 2008, S. 262).

Durch diese Methodik der Vergabe von mittleren Rängen kann es sehr leicht zu nicht ganzzahligen Rangsummen kommen, die in der theoretischen Teststatistik nicht vorgesehen sind. Dies führt dazu, dass kein exakter p-Wert mehr berechnet werden kann. Für diese Arbeit stellt dies beim Wilcoxon-Vorzeichen-Rang-Test zwar kein größeres Problem dar, da nur von Bedeutung ist, ob die Nullhypothese überhaupt verworfen wird, eine der geläufigsten Möglichkeiten; einen p-Wert dennoch zu berechnen, soll an dieser Stelle vorgestellt und im späteren Verlauf der Arbeit auch angewandt werden, die Normalverteilungsapproximation (Büning & Trenkler; 1994, S. 99).

Der Erwartungswert der Teststatistik ist gegeben als $E(W^+) = \frac{n(n+1)}{4}$ und die Varianz als $Var(W^+) = \frac{n(n+1)(2n+1)}{24}$ und diese Parameter können nun in eine Normalverteilung eingesetzt werden, also $W^+_{approx} \sim N(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24})$ (Fahrmeir et al.; 2007, S. 443). Diese Approximation an die wahre Verteilung funktioniert bereits bei einem relativ geringen Stichprobenumfang sehr gut, in der Literatur wird meist ein Wert von $n > 20$ angegeben, bei dem die Approximation eingesetzt werden kann (Fahrmeir et al.; 2007, S. 443). Abbildung 1 zeigt die Verteilung der Teststatistik bei einem Stichprobenumfang von nur $n = 10$ und die dazugehörige Normalverteilungsapproximation. Bereits hier kann man sehen, dass die Anpassung erstaunlich genau ist.

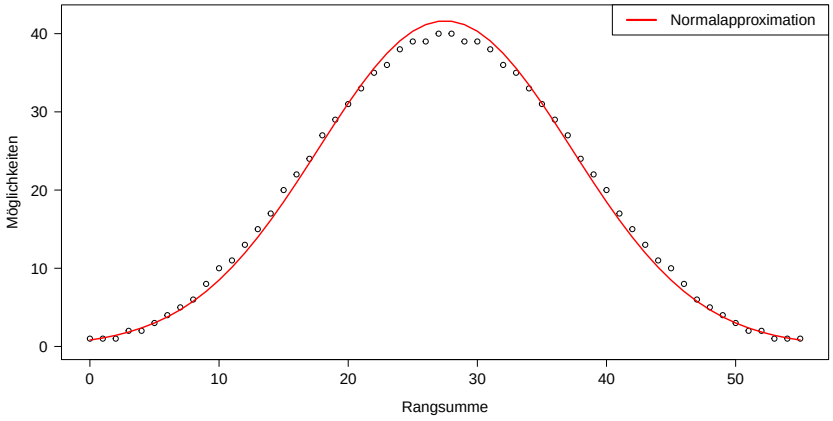

Abbildung 1: Verteilung der Teststatistik W^+ bei $n = 10$

Für den Fall, dass sich unter den Daten Ties befinden, bleibt zwar der Erwartungswert unberührt, die Varianz wird jedoch kleiner und muss mit einem Korrekturfaktor folgender Gestalt versehen werden, der von der ursprünglichen Varianz subtrahiert wird: (Büning & Trenkler; 1994, S. 99)

$$\sum_{i=1}^{n} \frac{t_i^3 - t_i}{48}$$

t_i bezeichnet die Anzahl jeder unterschiedlichen Absolutdifferenz $|D_i|$. Wenn keine Bindungen vorliegen, ist also jedes t_i gleich 1, der Korrekturfaktor gleich 0 um somit gilt $\frac{n(n+1)(2n+1)}{24} = \frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{n} \frac{t_i^3 - t_i}{48}$.

Zudem wird eine Stetigkeitskorrektur vorgenommen und 0.5 vom Erwartungswert subtrahiert (Büning & Trenkler; 1994, S. 35), wie dies beispielsweise auch standardgemäß im Programm *R* (R Core Team; 2013) angewandt wird. Der resultierende p-Wert des Wilcoxon-Tests mit Normalverteilungsapproximation ergibt sich dann also als die Verteilungsfunktion der Normalverteilung mit Erwartungswert $\frac{n(n+1)}{4} - 0.5$ und Varianz $\frac{n(n+1)(2n+1)}{24} - \sum_{i=1}^{n} \frac{t_i^3 - t_i}{48}$ an der Stelle W^+.

2.2.3 t-Test

Der parametrische t-Test im Einstichprobenfall kann sich mit der selben Testproblematik beschäftigen. Hier wird getestet, ob das arithmetische Mittel einer Stichprobe X signifikant kleiner als der hypothetische Wert μ_0 ist. Mit in die Berechnung der Teststatistik, im Folgenden T genannt, geht das arithmetische Mittel $\bar{X}$ und die Standardabweichung S der Stichprobe. Die Formel für T lautet folgendermaßen (Fahrmeir et al.; 2007, S. 437)

$$T = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$$

mit $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ und $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{(n-1)}$ und $S = \sqrt{S^2}$.

Diese Teststatistik ist in der Nullhypothese t-verteilt mit $n-1$ Freiheitsgraden. Die Alternativhypothese $H1 : \mu < \mu_0$ wird angenommen, falls T kleiner als das α-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden ($t_\alpha(n-1) = -t_{1-\alpha}(n-1)$) ausfällt. Wenn der Umfang der Stichprobe mindestens $n = 30$ beträgt, so können diese Quantile durch diejenigen der Standardnormalverteilung, $N(0,1)$, ersetzt werden, da sich die t-Verteilung mit zunehmendem Stichprobenumfang mehr und mehr an diese anpasst (Fahrmeir et al.; 2007, S. 437). Die an den t-Test gebundenen Voraussetzungen sind die unabhängigen und identisch normalverteilten Daten, es sei denn, der Stichprobenumfang ist größer als 30 (Fahrmeir et al.; 2007, S. 437).

2.3 Testgüte

Bevor überhaupt die Güte eines Tests definiert wird, stellt sich zunächst die Frage, wann ein Test überhaupt als 'gut' anzusehen ist. Wenn die Fehlerwahrscheinlichkeiten des Tests möglichst niedrig sind, scheint eine plausible Antwort darauf zu sein. Bei Testentscheidungen können zweierlei Arten von Fehlern entstehen, der Fehler erster und der Fehler zweiter Art.

Der Fehler erster Art meint, dass sich ein Test für die Alternativhypothese entscheidet, obwohl in Wirklichkeit die Nullhypothese gültig ist. Wenn die

Alternativhypothese zutrifft, ein Test dies jedoch nicht erkennt, so ist das ein Fehler zweiter Art (Fahrmeir et al.; 2007, S. 415 & 416). In der in dieser Arbeit vorliegenden Testproblematik mit $H_1 : \mu < \mu_0$ sind die möglichen Fehler hier zusammengefasst:

	Wahr: $\mu \geq \mu_0$ (H_0)	Wahr: $\mu < \mu_0$ (H_1)
Testentscheidung für H_0	Richtige Entscheidung	Fehler zweiter Art
Testentscheidung für H_1	Fehler erster Art	Richtige Entscheidung

Tabelle 1: Übersicht über korrekte und falsche Testentscheidungen

Wenn sich korrekterweise für H_0 entschieden wird, nennt man das auch 'Spezifität', die Erkennung einer falschen Nullhypothese als solche wird auch 'Sensitivität', 'Power', 'Trennschärfe' oder 'Güte' genannt. Die Gütefunktion gibt nun zu jedem Wert für μ die Wahrscheinlichkeit dafür aus, dass die Nullhypothese verworfen wird. Die Werte, die diese Funktion annimmt, hängen dann immer unmittelbar mit den Wahrscheinlichkeiten der Fehler erster und zweiter Art zusammen. Ist die Gütefunktion an einem Punkt, der in Wirklichkeit im H_0-Bereich ist, so gilt, dass sie gleich dem Fehler erster Art ist. Befindet sie sich an einem Punkt im Ablehnungsbereich, so ist die Gütefunktion genau die Gegenwahrscheinlichkeit zum Fehler zweiter Art (Fahrmeir et al.; 2007, S. 421 & 422). Eine ideale Gütefunktion – die in der Praxis nicht existiert – wäre im Bereich der Nullhypothese gleich 0 und im Bereich der Alternativhypothese gleich 1 (Fahrmeir et al.; 2007, S. 421). Da dies in der Realität nicht möglich ist, ist ein Test wünschenswert, der bereits bei einem wahren, knapp unter dem hypothetischen Wert μ_0 liegenden μ eine hohe Güte erzielt.

2.4 Verwendung randomisierter Tests

Wie bereits deutlich gemacht, sind die Teststatistiken des Vorzeichen- und des Wilcoxon-Vorzeichen-Rang-Tests diskret, sie können nur eine bestimmte Anzahl an unterschiedlichen Werten annehmen. Daraus resultiert wieder, dass nur ebenso viele p-Werte realisiert werden können, was bedeutet, dass grundsätzlich nicht auf jedem Niveau von exakt α getestet werden kann. Der Verständlichkeit halber sei an dieser Stelle ein konkretes Beispiel aufgeführt.

Sei der Umfang einer Stichprobe X gleich $n = 10$. Für den Test der Alternativhypothese $H_1 : \mu < \mu_0$ wird der Vorzeichentest herangezogen, als Teststatistik erhält man wie bekannt die Anzahl der positiven Differenzen $D_i = x_i - \mu_0$. Möglichkeiten für diese Teststatistik A gibt es dann genau 11, nämlich $\{0; 1; 2; ...; 10\}$ und damit als mögliche Testergebnisse die p-Werte $(\sum_{i=0}^{k} \binom{10}{i}) \cdot 0.5^{10}, k = \{0; 1; 2; ...; 10\}$, die in der folgenden Tabelle aufgeführt sind:

k	$\mathbb{P}(A \leq k)$	k	$\mathbb{P}(A \leq k)$
0	0.0010	6	0.8281
1	0.0107	7	0.9453
2	0.0547	8	0.9893
3	0.1719	9	0.9990
4	0.3770	10	1.0000
5	0.6230		

Tabelle 2: Sämtliche realisierbare p-Werte des Vorzeichentests bei $n = 10$

Wenn nun zum Beispiel auf einem Niveau von exakt 5% getestet werden soll, sieht das problematisch aus. Der größte mögliche p-Wert, der kleiner als 5% ist, beträgt 0.0107 (für $k = 1$) und der darauf folgende, kleinste mögliche p-Wert, der größer als 5% ist, beträgt 0.0547 (für $k = 2$). Als 5%-Quantil wird nun 2 genannt, doch ein exaktes Quantil für ebendiese 5% ist 2 nicht. Werden alle Tests mit einem p-Wert von unter 0.05 abgelehnt, wird in Wirklichkeit auf einem Niveau von 0.0107 getestet, was einen deutlich konservativeren Test bedeuten würde.

Grundsätzlich wäre es nun möglich, schlicht auf dem Niveau $\alpha = 0.0547$ zu testen statt 0.05. Da die Teststatistik des Wilcoxon-Tests allerdings ebenso nur begrenzt viele (nämlich $\frac{n(n+1)}{2} + 1$) und zudem zum Vorzeichentest unterschiedliche p-Werte annehmen kann, führt dies nicht zum Ziel.

An dieser Stelle sei nun die Möglichkeit genannt, wie in dieser Arbeit dennoch dieses exakte α-Testniveau erreicht werden kann, nämlich die der Randomisierung der (nonparametrischen) Tests.

Sei k_α das α-Quantil einer Teststatistik, also der Wert, der zum niedrigsten p-Wert größer als α führt, und $\phi(T)$ die Wahrscheinlichkeit, eine Nullhypothese bei resultierender Teststatistik T abzulehnen, so sähe ein nicht randomisierter Test beispielsweise folgendermaßen aus (Kauermann & Hothorn; 2014, S. 73)

$$\phi(T) = \begin{cases} 1, & T < k_\alpha \\ 0, & T > k_\alpha \end{cases}$$

Die Nullhypothese wird für $T < k_\alpha$ also sicher abgelehnt und für $T > k_\alpha$ sicher beibehalten. Außer Acht gelassen ist hierbei jedoch die Möglichkeit für $T = k_\alpha$, was bei einem randomisierten Test nicht der Fall ist.

Ein Signifikanzniveau von α bedeutet, dass die Wahrscheinlichkeit für den Fehler erster Art maximal α betragen darf. Da der Fehler erster Art in der in dieser Arbeit besprochenen Testsituation maximal für $\mu = \mu_0$ ist, muss also die Gütefunktion an diesem Punkt genau α betragen. Dies wird durch einen Parameter γ erreicht, der $\phi(T)$ folgendermaßen ergänzt (Kauermann & Hothorn; 2014, S. 73)

$$\phi(T) = \begin{cases} 1, & T < k_\alpha \\ \gamma, & T = k_\alpha \\ 0, & T > k_\alpha \end{cases}$$

γ ist dabei eine Zahl $\in \mathbb{R}$ und $\in [0;1]$ und gibt die Wahrscheinlichkeit an, H_0 zu verwerfen, wenn die Teststatistik T genau k_α entspricht. Diese Wahrscheinlichkeit muss nun wie erwähnt so bestimmt werden, dass die Gütefunktion an der Stelle $\mu = \mu_0$ den Wert α annimmt (Kauermann & Hothorn; 2014, S. 73). Dies wird folgendermaßen realisiert:

$$\begin{aligned} G(\mu_0) = \mathbb{P}(T < k_\alpha) + \gamma \cdot \mathbb{P}(T = k_\alpha) &\stackrel{!}{=} \alpha \\ \mathbb{P}(T \leq k_\alpha - 1) + \gamma \cdot \mathbb{P}(T = k_\alpha) &\stackrel{!}{=} \alpha \\ F(k_\alpha - 1) + \gamma \cdot f(k_\alpha) &\stackrel{!}{=} \alpha \\ \rightarrow \gamma &\stackrel{!}{=} \frac{\alpha - F(k_\alpha - 1)}{f(k_\alpha)} \end{aligned}$$

$F(x)$ ist hierbei die Verteilungsfunktion einer Teststatistik an einem Punkt x, $f(x)$ bezeichne die Dichtefunktion am entsprechenden Punkt.

Für das vorherige Beispiel (Vorzeichentest, $n = 10$, $\alpha = 0.05$ $k_\alpha = 2$) ergäbe sich γ zu:

$$\gamma \stackrel{!}{=} \frac{0.05 - 0.0107}{0.0547 - 0.0107} \approx 0.89$$

Sollte die Teststatistik also genau 2 positive Differenzen D_i ergeben, so wird per Zufall entschieden, ob H_0 verworfen wird oder nicht, wobei die Wahrscheinlichkeit für das Verwerfen etwa 89% beträgt. Grundsätzlich ist zu sagen, dass die Wahrscheinlichkeit zugunsten des Verwerfens der Nullhypothese umso höher ist, je näher die Verteilungsfunktion der Teststatistik im Punkt k_α tatsächlich bei α liegt, was aus der vorangehenden Formel leicht zu erkennen ist.

3 Simulation der Gütefunktionen

3.1 Aufbau

Um eine Gütefunktion zu simulieren, wird folgendermaßen vorgegangen: Zunächst müssen zu den verschiedenen Tests, deren Güte simuliert werden soll, Funktionen geschrieben werden. Diese Funktionen werden daraufhin so konzipiert, dass sie lediglich ausgeben, ob die Nullhypothese auf dem Niveau α abgelehnt wird oder nicht; α kann nach Belieben gewählt werden.

Die Gütefunktion gibt, wie bereits definiert, diejenige Wahrscheinlichkeit dafür an, dass die Nullhypothese abgelehnt wird, wenn ein bestimmter, wahrer Wert für μ gegeben ist. Da in der Realität der tatsächliche Wert jedoch nicht bekannt ist, werden nun Daten mit den verschiedensten Werten von μ simuliert. Es liegt die Alternativhypothese $H_1 : \mu < \mu_0$ vor, deswegen ist es trivial zu sagen, dass je kleiner der Mittelwert μ der simulierten Daten ist, desto höher tendenziell die Wahrscheinlichkeit, dass die Nullhypothese abgelehnt wird, also desto höher die Gütefunktion. Die Güte lässt sich nun punktweise schätzen, indem mit Zufallszahlen ein Test häufig durchführt und dabei der Anteil derjenigen Tests, die die Nullhypothese ablehnen, berechnet wird. Wenn $\mu = \mu_0$ gilt, sollte dieser Anteil im Schnitt genau α betragen. Denn die Nullhypothese ist damit gerade gültig und wenn sie trotzdem verworfen wird, entspricht das genau dem Fehler erster Art.

Durchgeführt wurden sämtliche Simulationen mit der Statistik-Software *R* (R Core Team; 2013).

3.2 Festlegungen

An dieser Stelle sei zusammengefasst, welche Tests bzw. vor allem welche Varianten der nonparametrischen Tests anhand der Simulationen näher betrachtet werden. Außerdem werden einige Größen festgelegt, die sich durch die ganze weitere Arbeit ziehen.

- Für den Vorzeichentest werden die zwei bereits in Abschnitt 2.2.1 erwähnten Methoden für den Umgang mit Nulldifferenzen betrachtet. Die Variante, die Nulldifferenzen außer Acht lässt und die Stichprobe um die entsprechenden Werte kürzt, und die Variante, die diesen Werten zufällige Vorzeichen zuweist. In beiden Fällen wird randomisiert, falls die sich ergebende Teststatistik dem 0.05-Quantil entspricht. Werden beide Varianten herangezogen, so ist in den Grafiken die Variante, die den Stichprobenumfang um die Nulldifferenzen reduziert als 'Vorzeichentest 1' und die Variante, die bei Nulldifferenzen zufällige Vorzeichen vergibt, als 'Vorzeichentest 2' gekennzeichnet.

- Für den Wilcoxon-Test gibt es aufgrund der zusätzlichen Möglichkeit von Bindungen zahlreiche Kombinationsmöglichkeiten, wie ein Test mit diesen

und mit Nulldifferenzen umgeht. Betrachtet wird ein Wilcoxon-Vorzeichen-Rang-Test, der Nulldifferenzen außer Acht lässt, gemittelte Ränge bei Bindungen berechnet und die Normalapproximation durchführt. Zudem wird die Variante angewandt, in der Nulldifferenzen die kleinsten Ränge zugewiesen, ebenso gemittelte Ränge bei Bindungen berechnet werden und die, sollte sich für die Teststatistik das 0.05-Quantil ergeben, randomisiert. In den Grafiken ist die Variante ohne Normalapproximation als 'Wilcoxon-Test 1' und die mit als 'Wilcoxon-Test 2' bezeichnet. Wenn nur ein Wilcoxon-Test in seiner Güte dargestellt wird, so entspricht das der Variante ohne Approximation.

- Zusätzlich wird der t-Test im Einstrichprobenfall betrachtet.
- Für das Signifikanzniveau wurde sich in der vorliegenden Arbeit in allen durchgeführten Simulationen für den gängigen Wert 0.05 entschieden. Das heißt, dass die Alternativhypothese mit höchstens 5% Irrtumswahrscheinlichkeit angenommen wird.
- Die Anzahl der durchgeführten Simulationsdurchläufe beträgt 10000. Diese vergleichsweise hohe Zahl erweist sich durchaus als sinnvoll. Bei kleinen Güteunterschieden zwischen den Tests überschneiden sich durch zufällige Abweichungen die Gütefunktionen, was bei 10000 Durchläufen kaum mehr der Fall ist. Hier kann auch bei kleinen Güteunterschieden meist eindeutig die höhere Güte ausgemacht werden. Außerdem werden bei 1000 Durchläufen nicht selten höhere Gütewerte bei kleineren Abweichungen zwischen μ und μ_0 ausgegeben; dies erscheint unplausibel und führt teils zu alles andere als glatten Kurven (siehe Abbildung 37 im Anhang).
- Was den Stichprobenumfang betrifft, wird dieser in allen Fällen auf $n = 20$ gesetzt. Häufig wird zusätzlich noch überprüft, wie sich die Gütefunktionen ändern, wenn dieser Stichprobenumfang auf $n = 10$ halbiert wird. An einer Stelle wird zudem ein Stichprobenumfang von 15 gewählt.

3.3 Einschub zu den verschiedenen Varianten des Wilcoxon-Tests

Es sei bereits an dieser Stelle aus gegebenem Anlass ein Ergebnis vorweggenommen: Güteunterschiede bei dem Wilcoxon-Vorzeichen-Rang-Test zwischen der Variante mit und ohne Normalapproximation sind bei einem Stichprobenumfang von 20 nicht mehr bemerkbar. Daher werden im Laufe der Arbeit beide Gütefunktionen nur dargestellt, wenn der Stichprobenumfang $n = 10$ ist. Da die Variante mit Normalapproximation die Nulldifferenzen von der Stichprobe eliminiert, werden auch im Falle potentieller Nulldifferenzen beide Gütefunktionen gezeigt. Abbildung 2 zeigt zu dieser Thematik die Differenzen der Güte der beiden Varianten des Wilcoxon-Vorzeichen-Rang-Tests. Negative Werte sprechen hierbei für die Überlegenheit der Variante ohne die Approximation. In dem beispielhaften Szenario sind in diesem Fall vom Stichprobenumfang abhängige

Gütefunktionen normalverteilter Daten mit einem Erwartungswert von $\mu = -0.5$ und der Testproblematik $H_1 : \mu < \mu_0$ simuliert worden; für die Anzahl der Simulationsdurchläufe wurde auch hier 10000 gewählt. Deutlich erkennbar ist die Unterlegenheit der Normalapproximation bei einem niedrigen Stichprobenumfang. Erst ab einem Stichprobenumfang von etwa 20 sind die Differenzen kleiner oder größer 0 etwa ausgeglichen.

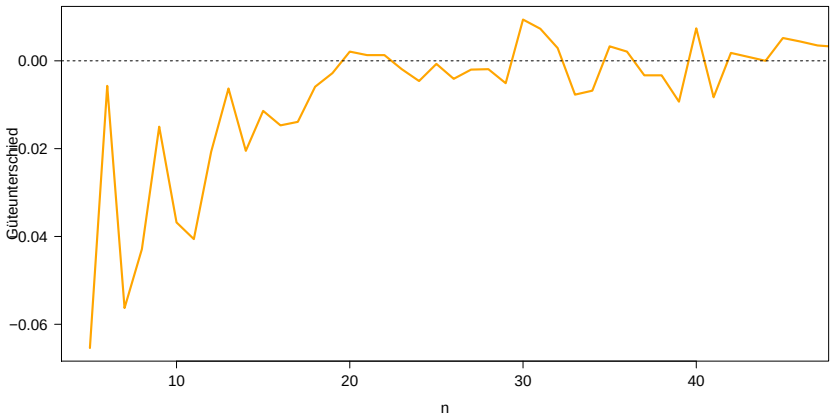

Abbildung 2: Differenzen zwischen der Güte der verschiedenen Varianten des Wilcoxon-Tests in Abhängigkeit des Stichprobenumfangs. Hierbei wurde die Güte der Variante ohne Normalapproximation von der mit Normalapproximation abgezogen.

Dieses Ergebnis deckt sich gut mit Literatur aus der Statistik, in der vielfach angegeben wird, dass die Teststatistik W^+ etwa ab einem Stichprobenumfang von 20 approximativ normalverteilt ist, siehe etwa Fahrmeir et al. (2007, S. 443).

3.4 Anwendung auf normalverteilte Daten

Normalverteilte Daten besitzen folgende Dichtefunktion:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}))$$

Der Erwartungswert liegt hierbei bei μ und die Varianz beträgt σ^2 (Abramowitz & Stegun; 1964, S. 930). Zunächst sollen die Tests auf normalverteilte Daten angewandt werden. Dabei werden diese jeweils mit Varianz 1 erzeugt mit unterschiedlichen negativen Mittelwerten bis hin zu 0 und darauf die Tests mit Alternativhypothese $\mu < \mu_0$ durchgeführt.

Im Hinblick auf die Literatur fallen die Ergebnisse nicht überraschend aus (Büning & Trenkler; 1994, S. 101). Diese zeigen - zu sehen in Abbildung 3 - dass der Vorzeichentest dem t-Test eindeutig unterlegen ist. Bei einem echten Mittelwert von kleiner als -1.2 und einem Stichprobenumfang von 20 erkennt der Vorzeichentest langsam zu 100%, dass der Mittelwert kleiner als 0 ist. Um eine Güte von 0.8 zu erreichen, bedarf es einer Abweichung von knapp 0.8 zwischen dem hypothetischen Wert 0 und dem wahren Mittelwert. Beim t-Test hingegen reicht hierfür eine Abweichung von knapp 0.6. Grundsätzlich verläuft die Gütefunktion des Vorzeichentests immer unter der des t-Tests, abgesehen trivialerweise von den Bereichen, an denen sie ineinanderlaufen, also an den Stellen, an der das wahre μ sehr klein ist und wo $\mu = \mu_0 = 0$ gilt. Der Wilcoxon-Vorzeichen-Rang-Test kann mit dem t-Test sehr gut mithalten. Seine Gütefunktion verläuft nur äußerst minimal unter der des t-Tests. Wenn μ kleiner als etwa -0.8 gewählt wird, laufen die Gütefunktionen bereits ineinander.

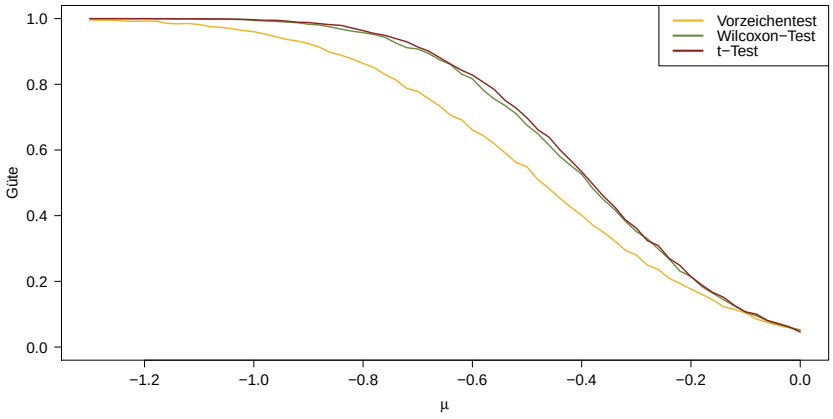

Abbildung 3: Gütefunktionen bei normalverteilten Daten und $n = 20$

Senkt man den Stichprobenumfang auf 10, machen sich vorher nicht oder nur schwer sichtbare Unterschiede zwischen den Tests bemerkbar (vgl. Abbildung 4). Gleich bleibt, dass der Vorzeichentest den anderen deutlich unterlegen ist. Auch hier gilt die minimale Überlegenheit des t-Tests dem Wilcoxon-Test gegenüber. Unter den Tests von Wilcoxon machen sich bei diesem niedrigen Stichprobenumfang nun Güteunterschiede bemerkbar. So scheint die Normalapproximation noch nicht genügend zu greifen, die Variante mit dieser Approximation unterliegt in der Güte der Variante ohne deutlich sichtbar.

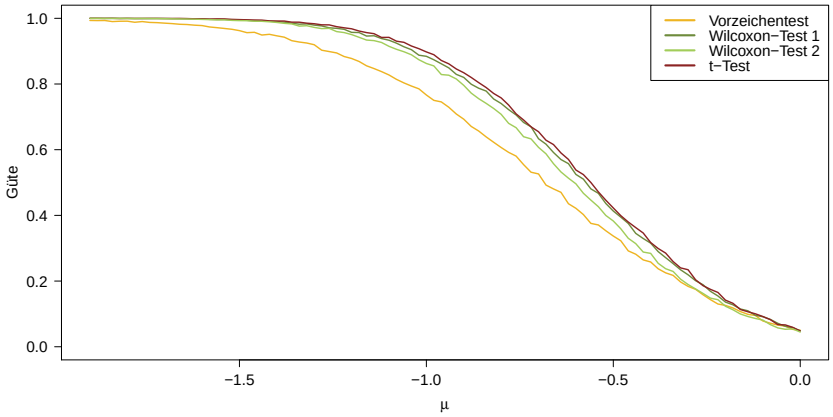

Abbildung 4: Gütefunktionen bei normalverteilten Daten und $n = 10$

3.5 Anwendung auf Daten anderer Verteilungen

3.5.1 Anwendung auf stetig gleichverteilte Daten

Stetig gleichverteilte Daten besitzen auf einem Intervall von a bis b die gleiche Wahrscheinlichkeitsdichte. Ihre Dichtefunktion ist somit gegeben zu:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{sonst} \end{cases}$$

Um die Daten dieser Verteilung besser mit den normalverteilten Daten im vorangegangenen Abschnitt vergleichen zu können, werden sie wieder auf $\mu = 0$ getestet und so generiert, dass die Varianz wieder gleich 1 ist.

Die Varianz von stetig gleichverteilten Daten beträgt $\frac{(b-a)^2}{12}$ (Abramowitz & Stegun; 1964, S. 930). Um eine Varianz von 1 zu erhalten, muss die Distanz zwischen der oberen und der unteren Grenze gleich $\sqrt{12}$ betragen. Damit werden Stichproben mit den Grenzen $[\mu - \frac{\sqrt{12}}{2}; \mu + \frac{\sqrt{12}}{2}]$ erstellt und für μ werden dabei wieder negative Werte bis hin zu 0 eingesetzt.

Bei einem Stichprobenumfang von 20 ist die Gütefunktion des t-Tests, wie in Abbildung 5 zu sehen ist, mit dem des gleichen Tests bei normalverteilten Daten zu vergleichen. Ab einer Abweichung von etwa 0.8 vom hypothethischen Wert 0 erreicht die Gütefunktion langsam einen Wert von 100%, bei einer Abweichung von 0.6 beträgt sie über 0.8. Deutlich sichtbar, doch noch relativ knapp darunter liegt die Gütefunktion des Wilcoxon-Tests. Der Unterschied zwischen dem t-Test und Wilcoxons Test fällt hier größer aus als bei den normalverteilten Daten. Unterschiede zwischen den beiden Vorgehensweisen der Wilcoxon-Tests sind nicht zu vermerken und daher nicht dargestellt – erst bei einer Senkung

des Stichprobenumfangs besitzt die Variante ohne Normalapproximation wieder eine höhere Güte. Der Vorzeichentest ist hier sehr deutlich abgeschlagen, noch stärker als bei normalverteilten Daten. Bei einer Abweichung von 0.6 werden nicht einmal die Hälfte aller Tests auf dem Niveau 5% abgelehnt, die Güte ist hier folglich unter 0.5. Erst ab einer Differenz zwischen μ_0 und μ von -1.3 führen langsam alle Testdurchführungen zu einer Ablehnung der Nullhypothese.

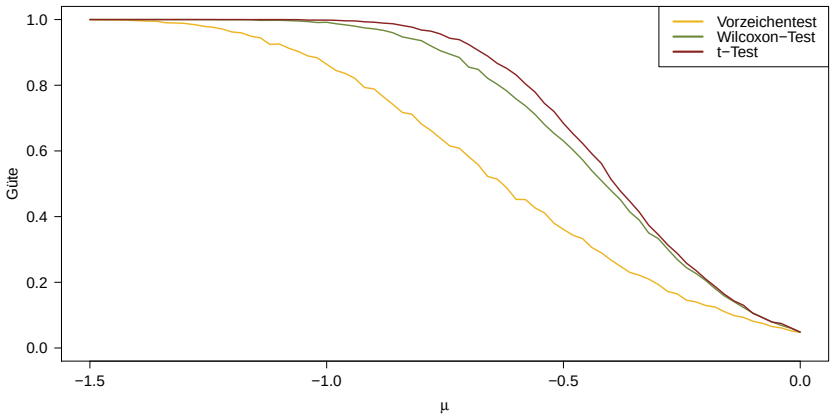

Abbildung 5: Gütefunktionen bei stetig gleichverteilten Daten und $n = 20$

Bei Senkung des Stichprobenumfangs werden – bis auf die Tatsache, dass sich nun Unterschiede zwischen den Varianten der Tests von Wilcoxon zeigen – keine auffälligen Ergebnisse beobachtet (siehe Abbildung 27 im Anhang).

3.5.2 Anwendung auf laplaceverteilte Daten

Die Laplaceverteilung ist eine stetige Verteilungfunktion, deren Dichtefunktion aussieht wie die Dichte einer Exponentialverteilung, an die senkrecht deren Spiegelung angefügt wird. Aus diesem Grund wird sie auch Doppelexponentialverteilung genannt. Ihre Dichtefunktion besitzt die folgende Gestalt:

$$f(x) = \frac{1}{2\sigma} exp(-\frac{|x-\mu|}{\sigma})$$

Der Erwartungswert laplaceverteilter Daten liegt bei μ, die Varianz ist gleich $2\sigma^2$ (Abramowitz & Stegun; 1964, S. 930). Um wieder Daten mit einer Varianz von 1 generieren zu können, muss $\sigma = \sqrt{0.5} \approx 0.7071$ gewählt werden.

Bei diesen Daten sieht man in Abbildung 6 bei einem Stichprobenumfang von 20, dass die nonparametrischen Tests eine höhere Güte aufweisen als der

t-Test. Am besten scheint der Wilcoxon-Vorzeichen-Rang-Test mit dieser Datenverteilung umgehen zu können, seine Gütefunktion liegt über denen der anderen Tests. Ab einem wahren Wert für μ von etwas weniger als -0.8 bewegt sich seine Gütefunktion auf 1 zu, was bei den anderen beiden Tests erst kurz vor -1.0 der Fall ist. Bei einer Differenz $\mu_0 - \mu$ kleiner als 0.5 überkreuzt sich seine Gütefunktion stark mit der des Vorzeichentests, sodass gesagt werden kann, dass diese hier die gleiche Güte aufweisen. Immer unterlegen ist den beiden Tests der t-Test. Grundsätzlich kann jedoch auch erwähnt werden, dass die Unterschiede zwischen allen Tests nicht riesig ausfallen. Geeigneter für diese Daten im Vergleich zu normalverteilten Daten scheinen die beiden nonparametrischen Tests zu sein, so fallen die Gütefunktionen bei gleicher Varianz bei laplaceverteilten Daten höher aus. Bei dem t-Test sind solche Unterschiede bei den verschiedenen Datenverteilungen nicht zu bemerken.

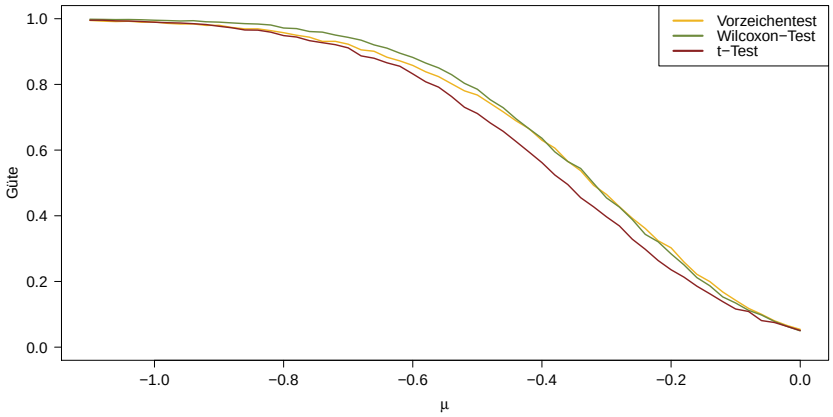

Abbildung 6: Gütefunktionen bei laplaceverteilten Daten und $n = 20$

Bei einem Stichprobenumfang von 10 ist der t-Test dem Vorzeichentest etwas überlegen, siehe Abbildung 28 im Anhang. Auch hier hat der Wilcoxon-Test die höchste Güte, zumindest in der Variante ohne Normalverteilungsapproximation.

Die Überlegenheit der nonparametrischen Tests bei laplaceverteilten Daten ist durch seine langen Tails begründet. Dies bedeutet, dass die Wahrscheinlichkeit für extremere Werte vergleichsweise hoch ist bei dieser Datenverteilung (siehe dazu auch Pagenkopf (1977, S. 83)).

3.5.3 Anwendung auf gemischt verteilte Daten

Folgten die bisher betrachteten Daten nur einer Verteilung, so wird nun betrachtet, wie sich die Gütefunktionen verhalten, wenn die Daten mehrerer Verteilungen folgen. Simuliert werden hierbei im Folgenden Daten, die mit einer gewissen

Wahrscheinlichkeit einer der drei bisher betrachteten Verteilungen folgen, also der Normal-, Laplace- oder der stetigen Gleichverteilung.

Zunächst wird untersucht, wie sich die Gütefunktionen verhalten, wenn die Daten mit gleicher Wahrscheinlichkeit der Normal- oder der Laplaceverteilung mit jeweils identischem Erwartungswert und identischer erwarteter Varianz folgen. Im Mittel ist dabei also auszugehen, dass zehn Werte normalverteilt und zehn weitere laplaceverteilt sind.

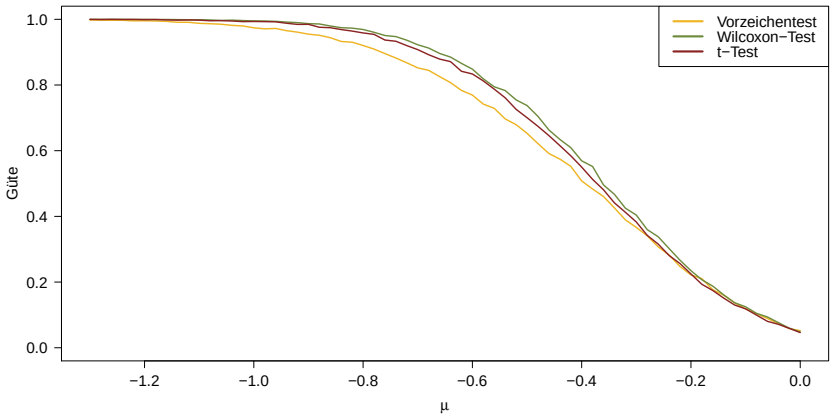

Abbildung 7: Gütefunktionen bei Daten, die mit gleicher Wahrscheinlichkeit entweder einer Normal- oder einer Laplaceverteilung mit gleicher Varianz folgen ($n = 20$)

Zu sehen ist in Abbildung 7, dass hier der Wilcoxon-Vorzeichen-Rang-Test mit leichtem Abstand zum t-Test grundsätzlich die höchste Güte aufweist. Der Vorzeichentest ist ein wenig abgeschlagen zu den anderen Tests, was seine Gütefunktion anbelangt, allerdings ist der Unterschied zwischen seiner Güte und der des t-Tests bei weitem nicht so groß wie bei pur normalverteilten Daten.

Nun werden die Daten so generiert, dass die Wahrscheinlichkeit für eine Laplaceverteilung nach wie vor 50% beträgt, die restlichen 50% fallen allerdings darauf, dass die Daten einer stetigen Gleichverteilung folgen. Wie zuvor werden diese Daten so generiert, dass im Mittel immer eine Varianz von 1 zu erwarten ist.

Es ergibt sich, dass erneut der Vorzeichentest beiden anderen Tests unterlegen ist, wenn auch wieder nicht allzu stark (Abbildung 8). Die Güteunterschiede zwischen Wilcoxons-Vorzeichen-Rang- und dem t-Test fallen hier extrem gering aus. Nur bei genauer Betrachtung sieht man eine ausgesprochen kleine Überlegenheit des t-Tests.

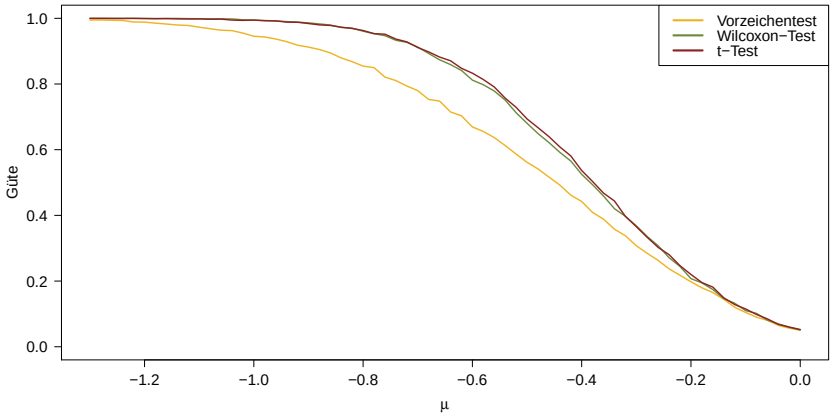

Abbildung 8: Gütefunktionen bei Daten, die mit gleicher Wahrscheinlichkeit entweder einer stetigen Gleich- oder einer Laplaceverteilung mit gleicher Varianz folgen ($n = 20$)

Die Gestalt der Gütefunktionen in den vorangegangenen Abbildungen 7 und 8 erscheint relativ intuitiv. So ist beispielsweise die Überlegenheit des Wilcoxon-Tests dem t-Test gegenüber bei laplaceverteilten Daten größer als seine Unterlegenheit bei normalverteilten Daten. Stammen die Daten mit gleicher Wahrscheinlichkeit aus einer der beiden Verteilungen, so verwundert es nicht, dass nun der Wilcoxon-Vorzeichen-Rang-Test hier eine leicht höhere Güte als der t-Test aufweist. Grundsätzlich besteht hier der Eindruck, dass die verschiedenen Verteilungen den gleichen Einfluss haben, wenn die Daten mit den gleichen Wahrscheinlichkeiten aus diesen Verteilungen stammen.

3.6 Anwendung auf problematische Daten

3.6.1 Anwendung auf tri- und bimodale Daten

In diesem Abschnitt werden Daten betrachtet, die an mehreren Stellen häufige Ausprägungen haben. Mit einer Wahrscheinlichkeit von ϵ sind die Daten nicht mit Erwartungswert um 0 verteilt, sondern entweder normalverteilt mit einem niedrigeren erwarteten Mittelwert oder einem höheren.

Die Daten folgen also entweder $N(\mu - \xi, 1)$ mit einer Wahrscheinlichkeit von $\frac{\epsilon}{2}$ oder $N(\mu, 1)$ mit einer Wahrscheinlichkeit von $1 - \epsilon$ oder aber $N(\mu + \xi, 1)$, die Wahrscheinlichkeit hierfür ist wieder $\frac{\epsilon}{2}$.

Je höher ϵ gewählt wird, umso mehr verläuft die Verteilung in eine bimodale Verteilung über, je niedriger, desto ähnlicher ist die Verteilung einer Normalverteilung. Nun werden für mehrere ξ und mehrere ϵ die Gütefunktionen der verschiedenen Tests beobachtet. Dadurch, dass der durchschnittliche Anteil $\frac{\epsilon}{2}$ für die Daten, die $N(\mu - \xi, 1)$- und die, die $N(\mu + \xi, 1)$-verteilt sind, gleich groß

ist und sie im Schnitt die gleiche Distanz, nämlich ξ, zu den $N(\mu, 1)$-verteilten Daten haben, ist eine symmetrische Verteilung dennoch gewährleistet und ganz gleich, wie μ gewählt wird, bleibt μ der Erwartungswert der erstellten Daten.

Für die Simulationen wird nun ein Stichprobenumfang von $n = 20$ und ξ gleich 10 gewählt, getestet wird wie gewohnt $H_1 : \mu < 0$ und die Daten werden mit entsprechenden Werten für μ generiert. Ist ϵ gleich 0, so entspricht die Verteilung wie erwähnt der einer ganz gewöhnlichen Normalverteilung und so bedarf dies keiner weiteren Betrachtung. Zunächst wird stattdessen ein Wert für ϵ von 10% gewählt. Bei $n = 20$ ist also zu erwarten, dass im Mittel genau ein Wert um $\mu - 10$ und ein Wert um $\mu + 10$ liegt und die restlichen 18 Werte um μ. Bei den Gütefunktionen zeigt sich ein deutliches Ergebnis: Der t-Test ist den nonparametrischen Tests eindeutig unterlegen (Abbildung 9). Auch wenn bezüglich der Güte im Vergleich zu normalverteilten Daten alle Tests einbüßen, so fällt der Unterschied der Gütefunktion zwischen diesen und den normalverteilten Daten beim t-Test doch enorm aus. Eine Güte von 0.8 erreicht der t-Test hier erst ab einer Differenz $\mu - \mu_0$ von betragsmäßig mehr als 2, für eine Güte von 100% bedarf es einer Differenz von zwischen -3 und -4. Eine Güte von 0.8 erlangen die beiden nonparametrischen Tests bereits bei einem Abstand zwischen dem wahren μ und dem hypothetischen μ_0 von weniger als 1. Auffällig ist, dass der Wilcoxon-Vorzeichen-Rang-Test die volle Güte erst bei μ knapp über -5 erreicht und zwischen diesem Wert und etwa -1 Gütewerte von über 0.9 erzielt. Eine Güte von knapp 100% erreicht der Vorzeichentest hier hingegen bereits zwischen $\mu = -1$ und $\mu = -2$.

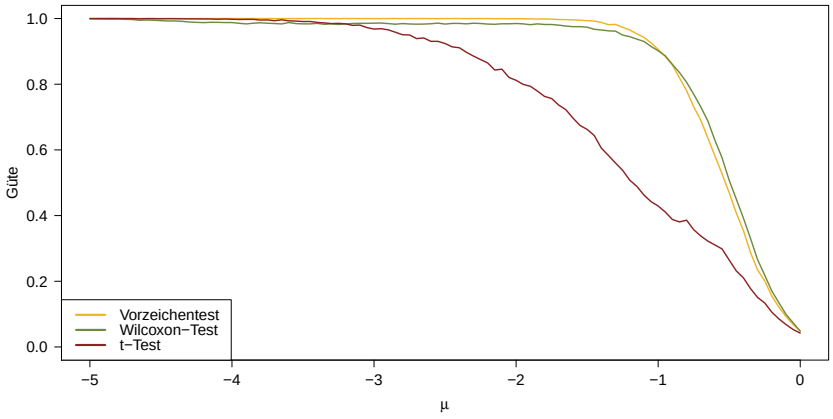

Abbildung 9: Gütefunktionen bei trimodalen Daten ($\epsilon = 10\%$, $n = 20$)

Woher der gravierende Unterschied zwischen dem t-Test und den nonparametrischen Tests an dieser Stelle rührt, lässt sich unschwer erklären. Man gehe der Einfachheit halber von dem erwarteten Fall aus, dass genau ein Wert unter

den Daten um $\mu - 10$ und genau ein Wert um $\mu + 10$ liegt. Welche Auswirkungen hat das auf die einzelnen Tests? Für den Vorzeichentest bedeutet dies, dass mit extrem hoher Sicherheit der Wert um $\mu + 10$ ein positives Vorzeichen annimmt (bei μ im betrachteten Wertebereich von -5 bis 0). Das 5%-Quantil des Vorzeichentests liegt bei $n = 20$ bei 6, was bedeutet, dass neben diesem 'determinierten' positiven Vorzeichen noch 4 weitere Vorzeichen positiv sein dürften, damit die Nullhypothese noch sicher abgelehnt wird. Diese 'Einschränkung' hat im Hinblick auf die Gütefunktionen offenbar noch recht geringe Auswirkungen. Der Vollständigkeit halber sei noch erwähnt, dass das Vorzeichen bei dem Wert um $\mu - 10$ bei dem betrachteten Wertebereich mit Sicherheit negativ sein dürfte. Beim Wilcoxon-Vorzeichen-Rang-Test sieht die Situation ähnlich aus. Es kann davon ausgegangen werden, dass die zwei größten Differenzen $D_i = |x_i - \mu_0|$ bei den Punkten um $\mu - \xi$ und $\mu + \xi$ entstehen. Der Wert um $\mu + \xi$ wird dementsprechend äußerst sicher zu dem Rang 20 oder 19 (bei sinkendem μ immer wahrscheinlicher 19) führen, was bei einem 5%-Quantil von 61 bei einer Stichprobe vom Umfang 20 bereits nicht unerheblich ist, jedoch bei weitem nicht alleine zur Beibehaltung von H_0 führt. Größere Auswirkungen als bei dem Vorzeichentest sind bezüglich der Güte und im Vergleich zu herkömmlichen normalverteilten Daten dennoch zu beobachten. Doch wie sieht es beim t-Test aus? Der t-Test lehnt im betrachteten Fall die Nullhypothese bekanntermaßen ab, wenn die Teststatistik $T = \frac{\bar{X} - \mu_0}{S}\sqrt{n}$ kleiner ausfällt als das Quantil der t-Verteilung mit $n - 1$ Freiheitsgraden. n ist hier 20, das entsprechende 5%-Quantil ≈ -1.7291 und μ_0 gleich 0. Die Stichprobenvarianz bei den normalverteilten Daten von vorhin beträgt 1 und hier beträgt sie im Schnitt 11, berechnet durch weitere Simulationen. Es kann also berechnet werden, wie groß $\bar{X}$ sein muss, damit die Nullhypothese bei durchschnittlicher Varianz überhaupt abgelehnt wird.

$$\begin{aligned} \frac{\bar{X} - 0}{\sqrt{11}}\sqrt{20} &\overset{!}{<} -1.7291 \\ \bar{X} &\overset{!}{<} \frac{-1.7291 \cdot \sqrt{11}}{\sqrt{20}} \\ \bar{X} &\overset{!}{<} -1.28 \end{aligned}$$

$\bar{X}$ müsste in diesem Fall also kleiner als -1.28 sein, damit die Nullhypothese abgelehnt werden kann. Im Fall einer Standardabweichung von 1 müsste $\bar{X}$ lediglich kleiner als etwa -0.39 sein. Der Grund für die drastische Verschlechterung der Güte des t-Tests ist also in der weitaus höheren Varianz der Daten bei nur zwei Ausreißern zu suchen.

Nun wird der Anteil an den Daten mit den Erwartungswerten von $\mu \pm \xi$ auf 25% erhöht. Zu sehen ist in Abbildung 10, dass sich die Gütefunktionen des t-Tests und des Vorzeichentests erwartungsgemäß weiter verschlechtern. Im Schnitt ist mit zwei oder drei positiven Vorzeichen zu rechnen, was den 'Spielraum' für den Vorzeichentest weiter senkt, was weitere positive Vorzeichen für

die Teststatistik betrifft, um trotzdem eine Ablehnung der Nullhypothese erreichen zu können. Auffällig bei diesem Verlauf ist lediglich, dass μ sehr klein ($\approx$ -9) gewählt werden muss, damit die Gütefunktion komplett 1 erreicht – eine Güte von über 0.98 erreicht der Vorzeichentest hingegen schnell.

Die Varianz für den t-Test erhöht sich auf durchschnittlich 26, was dementsprechend bedeutet, dass ein noch kleinerer Mittelwert der Daten benötigt wird, um die Nullhypothese $H_0 : \mu \geq 0$ ablehnen zu können.

Überraschender fällt hingegen der Verlauf der Gütefunktion des Wilcoxon-Tests aus. Die Funktion verläuft zunächst unter den beiden anderen Gütefunktionen, dann verharrt sie in einem Bereich von μ zwischen etwa -4 und -1.5 ungefähr auf 0.8, in diesem konstanten Verlauf übertrifft sie wieder die Gütefunktion des t-Tests und erst ab etwa -1.5 sinkt sie wieder, um sich dann langsam an die Gütefunktion des Vorzeichentests anzupassen.

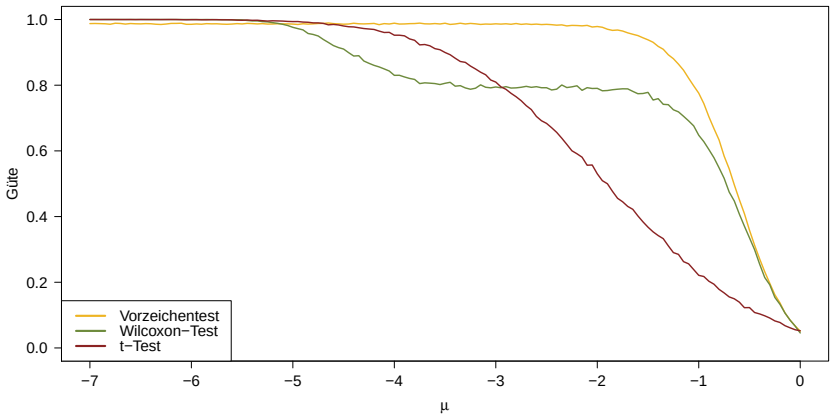

Abbildung 10: Gütefunktionen bei trimodalen Daten ($\epsilon = 25\%$, $n = 20$)

Als Erklärung für diesen ungewöhnlichen Verlauf kann Folgendes gesagt werden: Sei μ etwa -2. Erwartet werden dann bei $n = 20$ zwei oder drei Werte, die um -12, weitere zwei oder drei, die um 8 sowie etwa fünfzehn, die um -2 liegen. Der Einfachheit halber seien nun zwei Werte um -12, drei um 8 und fünfzehn um -2. Die Ränge -20 und -19 werden damit an die Werte um -12 vergeben, Rang 18, 17 und 16 an die Werte um 8. Alle restlichen Ränge sind als negativ zu erwarten und wenn positiv, dann sehr klein, was einen kleinen Rang bedeuten würde. Die Summe der positiven Ränge würde in diesem Fall 51 bedeuten und da dieser Wert unter dem 5%-Quantil liegt, wird H_0 in diesem plausiblen Fall abgelehnt. Nun sei μ gleich -3 und damit seien zwei Werte um -13, drei Werte um 7 und die restlichen fünfzehn um -3. An der Rangverteilung ändert sich im Vergleich zu $\mu = -2$ überhaupt nichts. Erst wenn μ so klein gewählt wird, dass die Wahrscheinlichkeit für $|x| \sim N(\mu, 1) > |x| \sim N(\mu + 10, 1)$ realistisch ist, steigt die Gütefunktion wieder an. Umgekehrt sinkt sie erst, wenn

μ betragsmäßig so klein ist, dass positive $x \sim N(\mu, 1)$ häufiger generiert werden.

Im Folgenden ist das Verhältnis zwischen den Daten mit Erwartungswert $\mu \pm \xi$ und denen mit Erwartungswert μ ausgeglichen bei 50% (siehe Abbildung 11). Zu beobachten ist, dass die Gütefunktion des t-Tests erneut auf einem niedrigeren Niveau verläuft, was bei einer Erhöhung der Varianz auf durchschnittlich 51 nur logisch ist. Der Verlauf der Gütefunktion des Wilcoxon-Tests ändert sich in seiner Form ebenso wenig, jedoch verharrt die Gütefunktion nicht mehr bei einem Wert von etwa 0.8, sondern weit darunter zwischen 0.4 und 0.5. Die Erklärung liegt darin, dass nun in diesem Bereich wieder mehr positive Ränge zu beobachten sind, die dann in ihrer Summe wesentlich häufiger das 0.05-Quantil der Teststatistikverteilung überschreiten als zuvor bei $\epsilon = 0.25$. Was sich grundlegend ändert, ist die Form der Gütefunktion des Vorzeichentests. Nun verharrt auch diese für gewisse Werte von μ auf einem Wert und zwar etwa zwischen $\mu = -2$ und $\mu = -8$ zwischen 0.7 und 0.8. Bei einem Wert für μ von beispielsweise -3 sind fünf positive Vorzeichen zu erwarten, nämlich genau die mit Erwartungswert $\mu + \xi = -3 + 10 = 7$. Die restlichen Daten haben entweder den Erwartungswert -3 oder -13 und somit sind dort negative Vorzeichen zu erwarten. Die erwarteten 5 positiven Vorzeichen würden bei einem 5%-Quantil von 6 noch zu einer Ablehnung der Nullhypothese führen, deshalb ist die Güte erwartungsgemäß hoch an dieser Stelle. Wird μ auf -6 geändert, so haben fünf Werte den Erwartungswert 4, weitere fünf den Erwartungswert -16 und zehn den Erwartungswert -6. Dies ändert an der erwarteten Anzahl an positiven Vorzeichen nichts und so ist die Form der Gütefunktion zu erklären. Erst ab dem Punkt, ab dem die Daten mit Erwartungswert $\mu + \xi$ eine höhere Wahrscheinlichkeit haben, negativ zu werden, steigt die Gütefunktion wieder an, was hier etwa bei $\mu = -8$ der Fall ist.

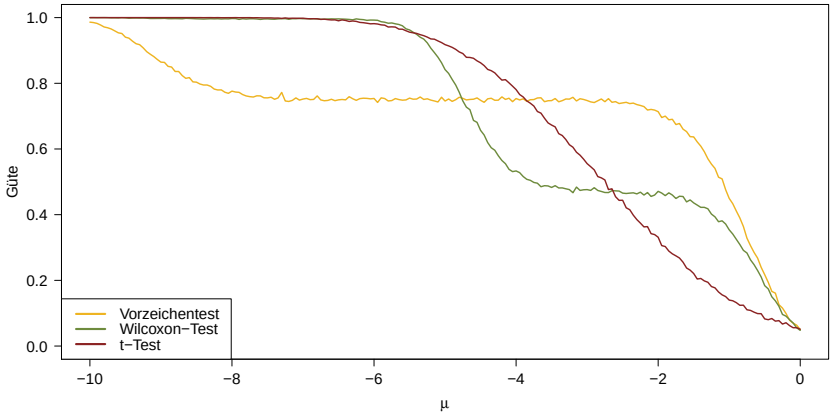

Abbildung 11: Gütefunktionen bei trimodalen Daten ($\epsilon = 50\%$, $n = 20$)

Während der t-Test bei $\epsilon = 0.75$ durch eine Varianz von durchschnittlich 76 weiter eine niedrigere Güte besitzt als zuvor, ändert sich auch die grundsätzliche Form der Gütefunktion des Vorzeichentests nicht, wie in Abbildung 12 ersichtlich. Etwa wieder in dem Bereich von μ zwischen -8 und -2 bleibt die Güte auf einem Wert, der hier nur noch etwa 0.3 beträgt. Zu erwarten sind in diesem Bereich im Schnitt 7.5 positive Vorzeichen, ein Testwert, der also zu groß für eine Ablehnung der Nullhypothese ist. Die Gütefunktion des Wilcoxon-Tests hat nun zwei Bereiche, in denen sie auf einem Wert stagniert. Etwa zwischen -8.5 und -6 auf ungefähr 0.9 und etwa zwischen -4 und -1.5 ungefähr auf 0.4. Im erstgenannten Bereich nehmen die Daten mit Erwartungswert $\mu + \xi$ im Schnitt betragsmäßig die kleinsten und die einzigen positiven Werte an, was zu den niedrigsten Rängen führt, die in ihrer Summe den Grenzwert meist nicht überschreiten. Im Bereich zwischen etwa -4 und -1.5 führen die Daten mit Erwartungswert μ im Schnitt zu den kleinsten Rängen (negative Ränge) und die mit Erwartungswert $\mu + \xi$ zu den positiven und zweitkleinsten, was dann vermehrt dazu führt, dass die entstehende positive Rangsumme höher ist als das 5%-Quantil.

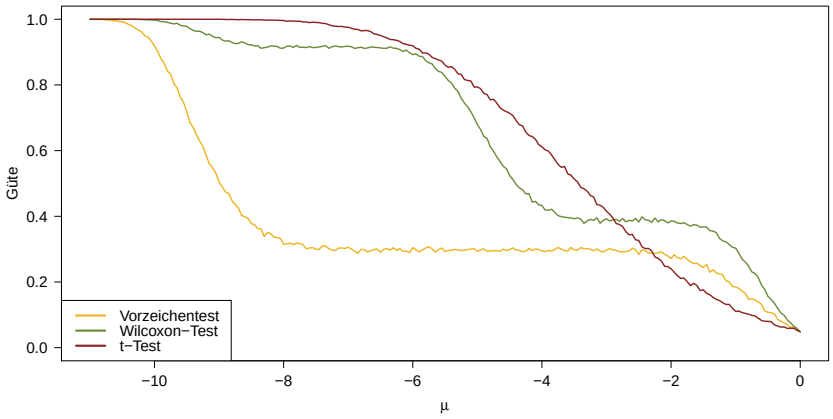

Abbildung 12: Gütefunktionen bei trimodalen Daten ($\epsilon = 75\%$, $n = 20$)

Bei $\epsilon = 100\%$ bestehen die Daten ausschließlich aus denen mit den Erwartungswerten -10 und 10 und eine rein bimodale Verteilung liegt vor. Die Varianz ist mit durchschnittlich 101 nun so hoch, dass μ bereits auf etwa -6 gesetzt werden muss, damit der t-Test eine Güte von ungefähr 0.8 erreicht (Abbildung 13). Zwischen etwa $\mu = -1.5$ und $\mu = -8.5$ ist die Gütefunktion des Wilcoxon-Tests konstant auf etwa 0.6, nur davor und danach steigt und sinkt sie. Das liegt daran, dass im Schnitt die Hälfte der Daten nun positiv und die Hälfte negativ ist und die negativen in diesem Bereich durchschnittlich die höheren zehn Ränge einnehmen. Zu erwarten ist also im Schnitt eine Rangsumme von $\sum_{i=1}^{10} i = 55$,

die noch – relativ knapp bei einem Grenzwert von 61 – zu einer Ablehnung der Nullhypothese führen würde. Der Vorzeichentest ist in diesem Fall deutlich ausgedrückt vollkommen unbrauchbar. Seine Güte ist bis etwa $\mu = -8$ konstant auf etwa auf $\alpha = 0.05$. In diesem Bereich sind positive und negative Vorzeichen im Verhältnis 1:1 zu erwarten und damit eine Teststatistik von 10, die zur Beibehaltung der Nullhypothese führt. Dass die Gütefunktion genau bei α liegt, hat den Grund, dass der Fehler erster Art so groß ist und wenn die Nullhypothese zutrifft (wovon der Test in diesem Bereich ausgeht) im Schnitt trotzdem ein Anteil von α aller durchgeführten Tests zur Ablehnung führen.

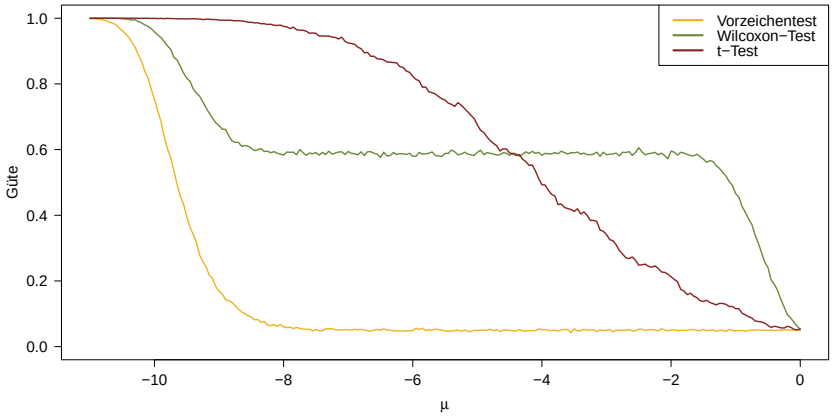

Abbildung 13: Gütefunktionen bei bimodalen Daten ($\epsilon = 100\%$, $n = 20$)

Betrachtet werden die Tests nun auch mit einem Wert für ξ von 3. Die Ergebnisse seien an dieser Stelle nur kurz angeschnitten. Bei $\epsilon = 0.1$ liegen die Gütefunktionen nahe beieinander, die des Wilcoxon-Tests liegt knapp über der des t-Tests, welche der des Vorzeichentests wiederum ein wenig überlegen ist. Liegt ϵ bei 25%, so überschneiden sich die Gütefunktionen immer wieder, wobei der Vorzeichentest für eine volle Güte den größten Abstand zwischen μ und μ_0 benötigt. Ab einem Wert für ϵ von 0.5 ist der t-Test den anderen Tests in Sachen Güte deutlicher überlegen, gerade der Vorzeichentest weist eine kleinere Güte auf. Dass die Güte von Tests auf einem konstanten Wert verharrt, ist allerdings nur ein wenig bei einer bimodalen Verteilung zu beobachten, hier bleibt die Gütefunktion des Wilcoxon-Tests kurz konstant bei $\epsilon = 100\%$. Die Güte des Vorzeichentests ist hier wieder länger lediglich bei einem Wert von α. Zusammengefasst kann also gesagt werden, dass sich bei einer Variation von ξ einiges ändert. Ist ξ mit 3 etwa verhältnismäßig klein, so werden die nonparametrischen Tests nur kaum systematisch beeinflusst, da sich die drei unterschiedlichen Verteilungen mehr untereinander überlappen, was bei einem Wert von $\xi = 10$ nur in extremen Ausnahmen der Fall ist. Der t-Test fällt nicht so sehr ab, da auch

die Varianz noch deutlich kleiner bleibt als bei $\xi = 10$. Die Grafiken zu den Gütefunktionen sind im Anhang (siehe Abschnitt A.2) zu finden.

3.6.2 Anwendung auf kontaminierte Daten

In diesem Abschnitt sollen die Daten nun sehr ähnlich zu Punkt 3.6.1 erstellt werden. Während in erwähntem Abschnitt Daten zu gewissen Wahrscheinlichkeiten verschiedenen Verteilungen folgen, so wird nun im Voraus festgesetzt, welcher Anteil der Daten welcher Verteilung folgt. Gleich bleibt, dass die Daten einer Normalverteilung mit Varianz 1 folgen und nur der Erwartungswert entweder μ, $\mu - \xi$ und $\mu + \xi$ betragen kann, wobei für ξ weiterhin der Wert 10 betrachtet wird. Der Anteil an Daten mit Erwartungswert $\mu - \xi$ und $\mu + \xi$ ist hierbei der gleiche. Nun seien zwei Beispiele hierfür aufgeführt:

Im Vergleich zu Abbildung 11 auf Seite 23 ist es in Abbildung 14 nicht mehr die Wahrscheinlichkeit $\epsilon = 0.5$, dass die Daten normalverteilt mit Erwartungswert $\mu \pm \xi$ statt μ sind – Stattdessen ist der Anteil $N(\mu + \xi, 1)$- und $N(\mu - \xi, 1)$-verteilter Daten nun fest bei jeweils $\epsilon^k = 0.25$. Deutlich sichtbar ist die Verschiedenheit beider Varianten. Die teilweise konstanten Verläufe der Gütefunktionen der nonparametrischen Tests sind hier nicht mehr zu beobachten. Zusätzlich gehen die Gütefunktionen von rechts schneller an 1 und von links statt auf $\alpha = 0.05$ direkt auf 0 zu. Zu erklären ist dies unter anderem damit, dass sowohl die Mittelwerte als auch die Rangsummen und Anzahl der positiven Vorzeichen der Daten wesentlich weniger streuen.

Die Gründe für die eben genannten Beobachtungen seien am Beispiel des Vorzeichentests genauer erklärt. In Abbildung 11 werden bei einem kleinen negativen Wert für μ 5 positive Vorzeichen (entstanden aus den Daten mit Erwartungswert $\mu + \xi$) und mit dem 5%-Quantil des Vorzeichentests von 6 die Ablehnung der Nullhypothese erwartet. Die Möglichkeit für mehr als diese fünf positiven Vorzeichen und damit die Beibehaltung der Nullhypothese ist allerdings weiterhin durchaus realistisch, was sich dadurch zeigt, dass die Gütefunktion lange auf etwa 0.8 verharrt. Ist nun aber festgelegt, dass genau fünf Werte $N(\mu + \xi, 1)$-verteilt sind, so ist die Möglichkeit für mehr als 5 positive Vorzeichen schnell äußerst gering. Bei knapp unter $\mu = -2$ verharrt die Gütefunktion des Vorzeichentests nun nicht mehr bei 0.8, sondern erreicht sofort 100%. Für eine Ablehnung der Nullhypothese müsste zu den nahezu sicheren 5 positiven Vorzeichen aus den $N(\mu + \xi, 1)$-verteilten Daten mindestens ein Wert der restlichen Daten positiv sein, was an der Stelle $\mu = -2$ und somit $N(-2, 1)$- und $N(-12, 1)$-verteilten Daten äußerst unwahrscheinlich ist. Und selbst, wenn 'nur' ein Wert positiv sein sollte und die Teststatistik mit $A = 6$ dem 5%-Quantil entsprechen würde, so würde das lediglich zur Randomisierung führen und dadurch nicht zwangsläufig zur Ablehnung der Nullhypothese.

Wieso die Güte bei $\mu = \mu_0 = 0$ statt auf 0.05 auf 0 zugeht, ist folgendermaßen zu erklären: Sei μ nun 0, so haben fünf Daten den Erwartungswert -10, zehn den Erwartungswert 0 und fünf den Erwartungswert 10. Sämtliche Daten mit betragsmäßigem Erwartungswert 10 werden durch die niedrige Varianz der Daten mit enormer Sicherheit das gleiche Vorzeichen haben wie ebendie-

ser Erwartungswert. Das bedeutet, dass an dieser Stelle alleine durch die nicht $N(\mu, 1)$-verteilten Daten 5 positive Vorzeichen stark zu erwarten sind. Die zehn Daten mit Erwartungswert 0 werden erwartungsgemäß zur Hälfte positiv und zur Hälfte negativ sein. Damit die Nullhypothese abgelehnt wird, dürfte dann höchstens einer der zehn um 0 verteilten Werte größer als 0 sein und selbst in diesem Fall würde noch randomisiert werden.

Gleichermaßen ist beim Wilcoxon-Vorzeichen-Rang-Test bis zu einem sehr kleinen μ zu erwarten, dass die $N(\mu + \xi, 1)$-verteilten Daten zur Rangsumme alleine mindestens den Wert $\sum_{i=11}^{15} = 65$ beitragen, was bereits das 5%-Quantil der Teststatistik von 61 überschreitet. Die Ränge von (-)16 bis (-)20 werden auf die $N(\mu - \xi)$-verteilten Daten fallen, da dort die betragsmäßig größten Differenzen zu erwarten sind.

Bei dem t-Test ist von einer hohen Varianz immer auszugehen. Da diese direkt mit der Testentscheidung zusammenhängt, führt bei betragsmäßig kleinem negativen μ keiner der Simulationsdurchläufe zur Ablehnung der Nullhypothese.

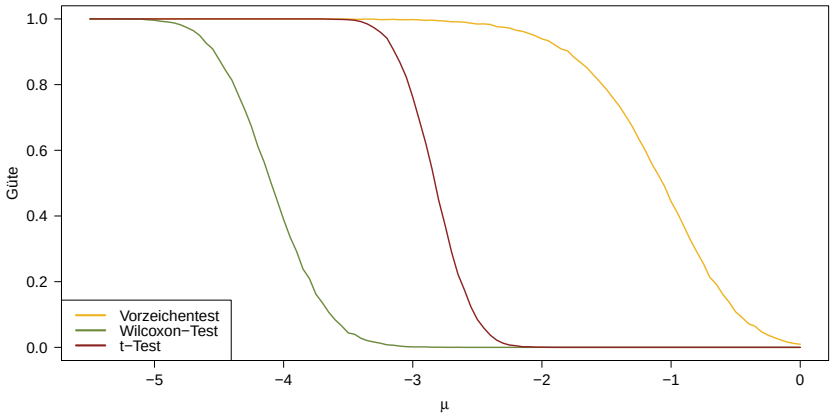

Abbildung 14: Gütefunktionen bei kontaminierten, trimodalen Daten ($\epsilon^k = 50\%$, $n = 20$, $\xi = 10$)

Nun wird ϵ^k auf 100% gesetzt, was bedeutet, dass zehn Werte den Erwartungswert $\mu - 10$ und die weiteren zehn den Erwartungswert $\mu + 10$ haben werden, was wieder einer bimodalen Verteilung entspricht. Somit können diese Daten direkt mit den bimodalen Daten aus Abschnitt 3.6.1 verglichen werden (deren Gütefunktionen siehe Abbildung 13) – Dort sind erwartungsgemäß auch jeweils zehn Daten aus den unterschiedlichen Verteilungen, was hier aber fest vorgegeben ist.

Zu gewinnen sind in Abbildung 15 wieder ähnliche Erkenntnisse wie bereits bei den vorangegangenen Daten mit $\epsilon^k = 0.5$. Erneut gibt es keine stellenweise konstant verlaufenden Gütefunktionen mehr zu sehen. Ebenso erreichen die

Gütefunktionen bereits im Bereich der Alternativhypothese Werte, die deutlich unter $\alpha = 0.05$ liegen, hier in allen Fällen komplett bei 0, es führen dort also sämtliche Simulationsdurchläufe in allen drei Tests zur Beibehaltung der Nullhypothese. An diesem Beispiel sei nun der Verlauf der Gütefunktion des Wilcoxon-Vorzeichen-Rang-Tests näher erläutert. Wo diese in Abbildung 13 bei abnehmendem μ zunächst auf einen Wert von etwa 0.6 steigt und auch dort verharrt, nimmt sie hier nun direkt 1 an. Zu erklären ist dies damit, dass zwar in beiden Fällen bei leicht negativen Werten für μ eine Rangsumme von $\sum_{i=1}^{10} = 55$ zu erwarten ist, was bei einem α-Quantil der Teststatistik von 61 zu einer Ablehnung der Nullhypothese führen würde; nun ist es im kontaminierten Fall jedoch so, dass mit viel höherer Wahrscheinlichkeit eben genau diese zehn positiven Ränge vorzufinden sein werden und hier beinahe kein Spielraum mehr besteht, der erhebliche Auswirkungen auf die Rangsumme hat. Im Fall der bimodalen Daten aus Abschnitt 3.6.1 kann es leicht vorkommen, dass mehr als zehn positive Ränge entstehen, was zwangsläufig zu einer höheren Rangsumme führt und damit die Nullhypothese häufiger beibehalten wird (da die Gütefunktion auf ungefähr 0.6 verharrt, in etwa 40% der Fälle). Die Gütefunktion des Vorzeichentests ändert sich bis auf die Tatsache, dass sie sehr lange auf einem Wert von 0 statt $\alpha = 0.05$ verharrt, kaum im Vergleich zu den bimodalen Daten aus Abschnitt 3.6.1.

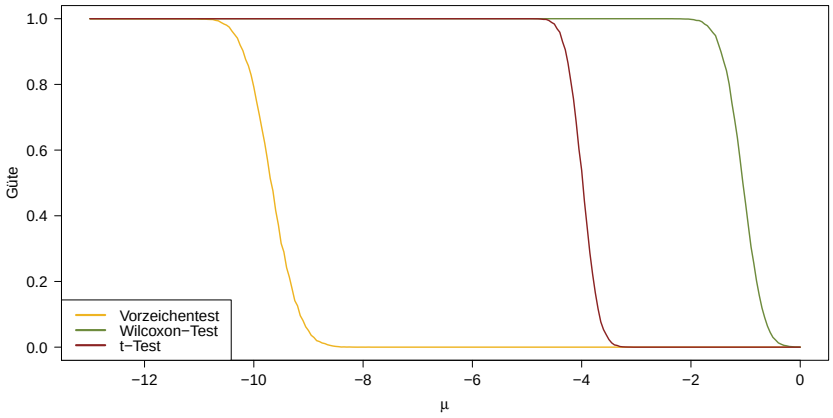

Abbildung 15: Gütefunktionen bei kontaminierten, bimodalen Daten ($\epsilon^k = 100\%$, $n = 20$, $\xi = 10$)

Zu dieser Art von Daten ist zusammengefasst zu sagen, dass sie sämtliche Tests systematisch beeinflussen, indem der Bereich zwischen 0 und 1 der Gütefunktionen sehr viel kleiner wird und dass nicht auf einem Niveau von α getestet werden kann. Letzteres ist skeptisch zu sehen, da die Tests teilweise selbst bei einer deutlichen Unterschreitung von 0 des wahren Werts für μ mit einer simu-

lierten Güte von 0 nicht in der Lage sind, die zutreffende Alternative $\mu < \mu_0$ zu erkennen.

3.6.3 Anwendung auf Daten mit großer Varianz

Da im Abschnitt 3.6.1 festgestellt wurde, dass der Grund für die stark niedrigere Güte des t-Tests bei Daten mit Ausreißern im Vergleich zu solchen ohne in der weitaus höheren Varianz begründet liegt, soll nun überprüft werden, wie sich die Tests verhalten, wenn Daten ohne Ausreißer getestet werden, die aber dennoch die gleiche Varianz besitzen wie die Daten vorhin mit Ausreißern. Im vorangegangenen Abschnitt lag die Varianz der Daten bei einem Ausreißeranteil von $\epsilon = 0.1$ bei durchschnittlich 11. Nun werden also zum Vergleich normalverteilte Daten mit Varianz 11 erstellt. Beim t-Test ist die gleiche Gütefunktion zu erwarten, da dessen Teststatistik nur die Parameter Mittelwert und Varianz beinhaltet und diese in der folgenden Simulation wie erwähnt identisch zu der Simulation in Abschnitt 3.6.1 sein werden. Die Frage, die sich stellt, ist, wie sehr die Gütefunktionen der nonparametrischen Tests beeinflusst werden, wenn zwar die erwartete Varianz gleich bleibt, deren Ausmaß allerdings nicht mehr zum Großteil durch Ausreißer zu begründen ist.

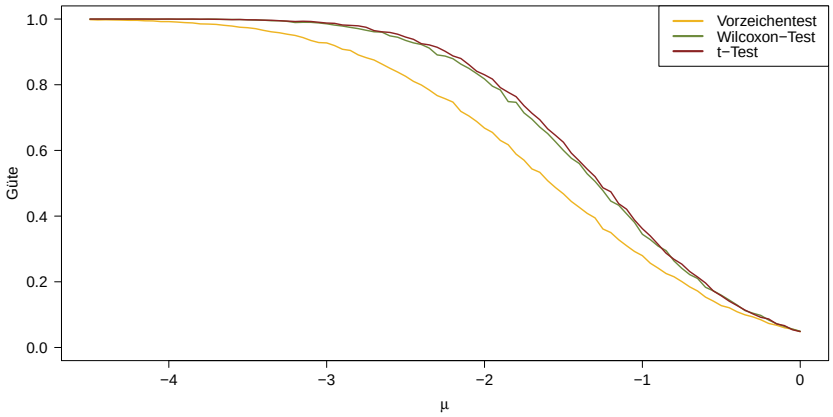

Abbildung 16: Gütefunktionen bei normalverteilten Daten mit Varianz 11 und $n = 20$

An den Gütefunktionen in den Abbildungen 16 und 17 ist zu erkennen, dass – wie erwartet – die des t-Tests verläuft wie bereits in Abschnitt 3.6.1. Die nonparametrischen Tests nehmen jedoch eine komplett andere Gestalt an. Wie bereits bei den normalverteilten Daten mit Varianz 1 aus Abschnitt 3.4 ist die Gütefunktion des Wilcoxon-Vorzeichen-Rang-Tests der des t-Tests nur minimal unterlegen. Der Vorzeichentest besitzt eine deutlich niedrigere Güte als die beiden anderen Tests. Auch bei normalverteilten Daten mit einer erwarteten

Varianz von 26 (entspricht der durchschnittlichen Varianz bei den trimodalen Daten mit $\xi = 10$ und $\epsilon = 0.25$) sind ähnliche Ergebnisse zu beobachten. Sind die Daten normalverteilt und man ändert lediglich deren Varianz, wirkt sich das also nur insofern aus, dass es für das Erreichen einer bestimmten Güte eine entsprechend größere oder kleinere Abweichung zwischen μ und μ_0 bedarf.

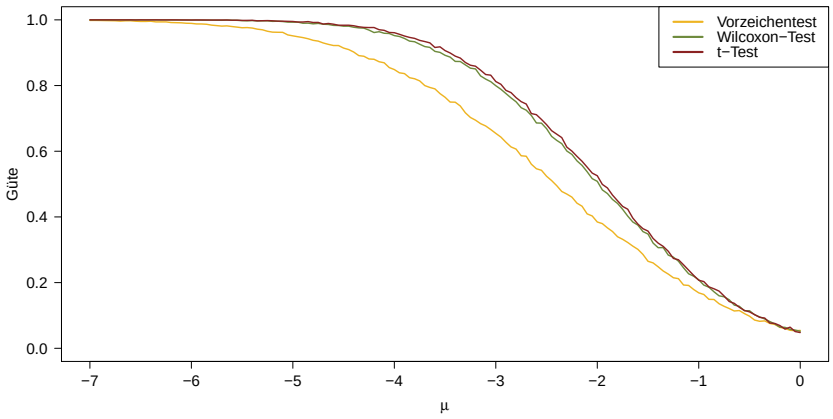

Abbildung 17: Gütefunktionen bei normalverteilten Daten mit Varianz 26 und $n = 20$

Die Überlegenheit der nonparametrischen Tests bei einem kleinen Anteil an Ausreißern ist also nicht auf die Varianz zurückzuführen. Wenige Ausreißer wirken sich auf deren Güte nur schwächer aus als auf den t-Test.

3.6.4 Anwendung auf gerundete Daten

In der Praxis kann mit den verfügbaren Messinstrumenten nie auf eine beliebige Genauigkeit gemessen werden. Irgendwann stoßt ein Messgerät an seine Grenzen und eine größere Präzision kann nicht erreicht werden. Dieser Thematik entsprechend werden in diesem Abschnitt nun Daten betrachtet, die nur auf wenige Kommastellen genau angegeben werden. Dazu werden zunächst wieder normalverteilte Daten mit Varianz 1 erzeugt – ganz im Sinne von Abschnitt 3.4. Diese Daten werden dann zunächst auf eine Nachkommastelle und danach auf eine ganze Zahl gerundet. Dadurch entstehen fast zwangsweise Bindungen und Nulldifferenzen und das Augenmerk liegt darauf, auszumachen, inwiefern diese die Tests mit der üblichen Problematik $H_1 : \mu < \mu_0 = 0$ beeinflussen.

Wenn auf eine Nachkommastelle gerundet wird, so lässt sich im Grunde kurz und knapp sagen: Es bestehen keine wirklich nennenswerten Unterschiede zu den Daten aus Kapitel 3.4. Die Gütefunktionen, die hier entstanden und in Abbildung 18 abgebildet sind, könnten ebenso aus jenen Daten entstanden sein,

die auf zahlreiche Nachkommastellen genau angegeben werden. Die Variante des Vorzeichentests, die Nulldifferenzen aus der Stichprobe entfernt, ist hier derjenigen mit zufälliger Rangvergabe minimal überlegen. Wird der Stichprobenumfang auf 10 gesetzt, so ergibt sich ein weiterer, kleiner Unterschied. Im direkten Vergleich zu den normalverteilten und ungerundeten Daten zeigt sich bei dieser Stichprobengröße hier ein geringerer Unterschied zwischen den Gütefunktionen der verschiedenen Vorgehensweisen der Tests von Wilcoxon (vgl. Abbildung 34 im Anhang).

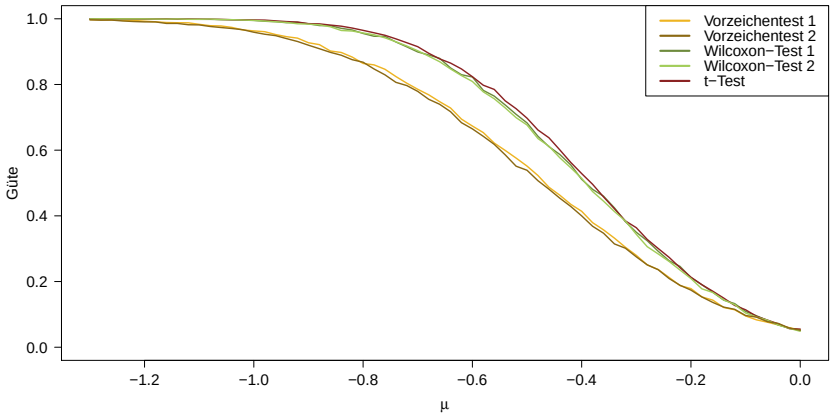

Abbildung 18: Gütefunktionen bei gerundeten Daten (eine Nachkommastelle) und $n = 20$

Um den Grund für die fast identischen Gütefunktionen auszumachen, wird nun getestet, wie stark in diesem Fall überhaupt Bindungen und Nulldifferenzen auftreten. Der Anteil der Bindungen wird erfasst, indem wie gehabt die Daten wie vorhin beschrieben generiert und der Anteil der unterschiedlichen Werte durch den Stichprobenumfang geteilt wird. Daraus resultiert, dass bei einem Stichprobenumfang von 20 der Anteil der verschiedenen Werte bei 100000 Durchgängen durchschnittlich knapp 0.78 beträgt. Bei einem Umfang von 10 ist dieser ein wenig höher, nämlich bei knapp über 0.88. Der Anteil von Bindungen ist also in beiden Fällen noch relativ gering.

Ähnlich wird dabei verfahren, wenn der Anteil an Nulldifferenzen simuliert werden soll. Hier muss beachtet werden, dass dieser tendentiell umso höher ist, je näher der wahre Mittelwert der Daten auch tatsächlich am hypothetischen Wert 0 ist. Hier ist der Anteil unabhängig vom Stichprobenumfang. Betrachtet wird er bei den wahren Mittelwerten von -1.3 bis 0 und steigt hier stetig an. Bei $\mu = -1.3$ beträgt der Anteil an Nulldifferenzen noch etwa 1.7%, bei $\mu = 0$ etwa 4%. Diese Anteile sind offenbar noch zu niedrig, als dass Unterschiede zwischen den verschiedenen Vorgehensweisen mit dem Umgang von Nulldifferenzen sicht-

bar zum Vorschein kommen.

Zwangsweise höher werden die Anteile von Bindungen und Nulldifferenzen, wenn auf eine ganze Zahl gerundet wird. Zunächst seien, analog zu den auf eine Kommastelle gerundeten Daten, diese Anteile genannt. Bei $n = 20$ ist der Anteil der unterschiedlichen Werte im Schnitt gerade noch bei knapp über 23%, was weniger als fünf unterschiedliche Werte ausmacht. Beträgt der Stichprobenumfang 10, so erhöht sich der Anteil auf knapp über 39%, was im Schnitt etwas weniger als vier unterschiedliche Werte bedeutet.

Im Hinblick auf die Anteile der Nulldifferenzen unter den Daten kann gesagt werden, dass diese sich etwa verzehnfachen, wenn zwischen Daten mit einer Nachkommastelle und Daten, die auf eine ganze Zahl gerundet sind, verglichen wird. Bei $\mu = -1.3$ beträgt der Anteil der Nulldifferenzen bereits weit über 17% und wächst auf bis knapp über 38% bei der Stelle $\mu = \mu_0 = 0$. Doch macht sich das auf die Gütefunktionen bemerkbar und wenn ja, wie?

Zunächst ist zu sagen, dass bei einem Stichprobenumfang von 20 alle Gütefunktionen minimal 'schlechter' erscheinen als bei den ungerundeten, normalverteilten Daten (Abbildung 19). Die Güte ist an den gleichen Stellen niedriger, für die gleiche Güte bedarf es einer größeren Differenz zwischen dem wahren μ und dem hypothetischen Wert $\mu_0 = 0$. Eine Ausnahme ist hierbei äußerst auffällig: Die Gütefunktion des Vorzeichentests, der Nulldifferenzen außer Acht lässt und aus der Stichprobe eliminiert ist höher als bei ungerundeten standardnormalverteilten Daten und kann mit Wilcoxons Tests nahezu mithalten, was überraschend ist. Die andere Variante, die bei Nulldifferenzen zufällig positive oder negative Vorzeichen verteilt, hat somit mit Abstand die niedrigste Güte. Zurückzuführen ist die überraschend hohe Güte einer Variante des Vorzeichentests wohl auf ebendiese Eliminierung der Nulldifferenzen. Wenn diese entfernt werden, bleiben, wenn μ kleiner wird, immer weniger positive Differenzen übrig, denn dafür müsste der ursprüngliche Wert mindestens 0.5 betragen, was bei sinkendem μ immer unwahrscheinlicher wird. Das Verhältnis zwischen den negativen und den positiven Differenzen wird so also immer größer und so werden immer mehr Nullhypothesen abgelehnt. Was davon zu halten ist, dass ein Test bei den eigentlich gleichen (normalverteilten) Daten, die jedoch wesentlich weniger Information enthalten, eine höhere Güte besitzt, gerade wenn er zwischen 17 und 40% aller Daten außer Acht lässt, sei dahingestellt. Der Eindruck, dass diese überlegene Güte mehr Schein als Sein ist, kommt jedoch wohl nicht zu Unrecht auf.

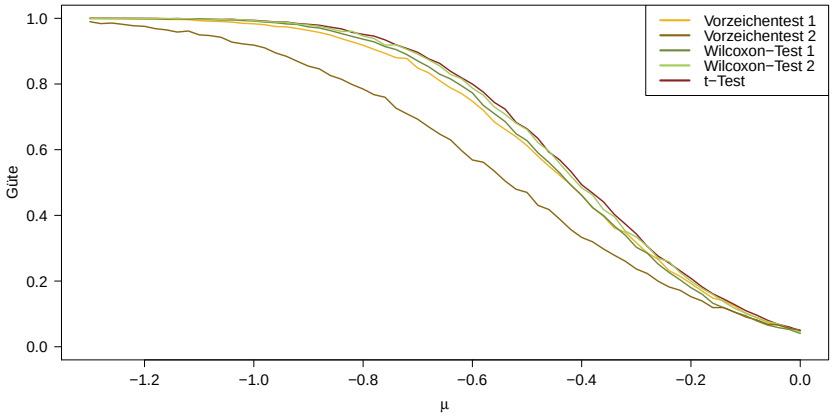

Abbildung 19: Gütefunktionen bei gerundeten Daten (ganze Zahl) und $n = 20$

Wilcoxons Tests können wieder gut mit dem t-Test mithalten, minimal sind die Unterschiede zwischen deren Gütefunktionen. Zu sagen ist hierbei, dass unter den Vorzeichen-Rang-Tests die Variante mit der Normalapproximation eine höhere Güte aufweist. Zu beachten ist, dass diese die Nulldifferenzen ebenso außer Acht lässt.

Bei einem Stichprobenumfang von 10 ist der t-Test den anderen Tests knapp, aber sichtbar überlegen. Unter den Tests von Wilcoxon weist die Variante mit Normalapproximation wie auch bei einem Stichprobenumfang von 20 die höhere Güte auf, hier jedoch nur minimal, was eine größere Überlegenheit des t-Tests den anderen Methoden gegenüber zur Folge hat. Hinsichtlich der Vorzeichentests ergeben sich keine nennenswerten neuen Erkenntnisse (Abbildung 35 im Anhang).

Was hier jedoch bei der Erstellung der Gütefunktionen zu einem grundlegenden Problem führt, ist Folgendes: Dadurch, dass die Daten auf eine ganze Zahl gerundet sind und der Stichprobenumfang so gering ist, kommt es bei 10000 Durchläufen durchaus dazu, dass hin und wieder alle Elemente den gleichen Wert annehmen. Problematisch ist dies, da dadurch eine Stichprobenvarianz von 0 entsteht und bei der Teststatistik des t-Tests, $T = \frac{\bar{X}-\mu_0}{S}\sqrt{n}$, durch die Wurzel ebendieser Stichprobenvarianz dividiert wird. In den Fällen, in denen diese 0 annimmt, kann argumentiert werden, dass der t-Test grundsätzlich nicht durchgeführt werden kann. Es wurde im Rahmen dieser Arbeit jedoch festgelegt, dass die Nullhypothese hierbei immer abgelehnt wird, wenn der Mittelwert $\bar{X}$ der Stichprobe den Wert μ_0 unterschreitet. Als Begründung hierfür sei bedacht, dass bei (in Relation zu $\bar{X}$) sehr kleinen – und damit nahe bei 0 gelegenen – Werten für S durchaus abgelehnt wird. Dennoch sei erwähnt, dass dieses Argument bei noch wesentlich kleineren Stichproben nicht zählt. Grafisch veranschaulicht ist das Auftreten einer Stichprobenvarianz von 0 bei diesen Daten und diesem

Wert für n in Abbildung 36 im Anhang.

Nun wird überprüft, ob und inwiefern sich der auffällige Unterschied zwischen den zwei Varianten des Vorzeichentests bei Rundung auf eine ganze Zahl verringert, wenn die Daten mit einer größeren Varianz generiert werden. Plausibel wäre dies, denn im Fall einer stärkeren Streuung der Daten sind mehr unterschiedliche Werte und weniger Nulldifferenzen zu erwarten. Als Standardabweichung der Daten werden im Folgenden Werte von 3 und 5 gewählt. Logisch dabei ist, dass die Gütefunktionen grundsätzlich niedriger ausfallen werden, doch dies ist nun nicht primär von Interesse.

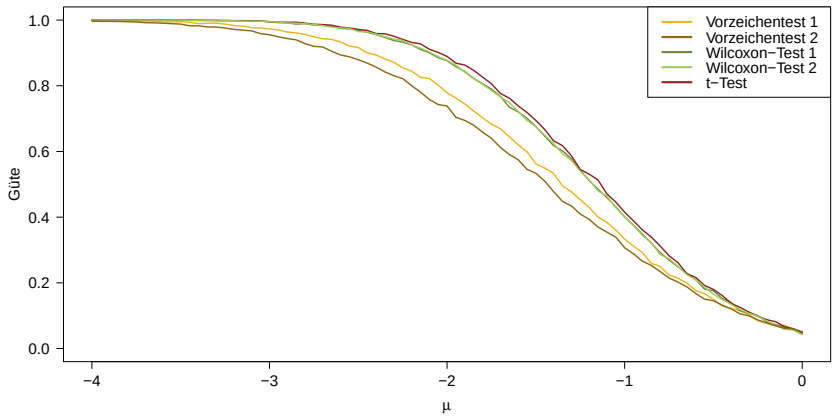

Abbildung 20: Gütefunktionen bei gerundeten Daten (ganze Zahl), S=3 und $n = 20$

Das Ergebnis der Beobachtungen ist, dass bei zunehmender Varianz die Unterschiede der Gütefunktionen der beiden Varianten geringer werden. Die Variante, die Nulldifferenzen aus der Stichprobe entfernt, kann bereits bei einer Standardabweichung von 3 (Abbildung 20) bei weitem nicht mehr mit dem t-Test und dem Wilcoxon-Vorzeichen-Rang-Test mithalten und so bestätigt sich der Eindruck, dass von der hohen Güte dieser Variante (vgl. Abbildung 19) nicht unbedingt viel zu halten ist. Trotzdem sei erwähnt, dass der Unterschied zwischen den beiden Varianten des Vorzeichentests auch bei einer Standardabweichung von 5 dennoch besteht (siehe Abbildung 21).

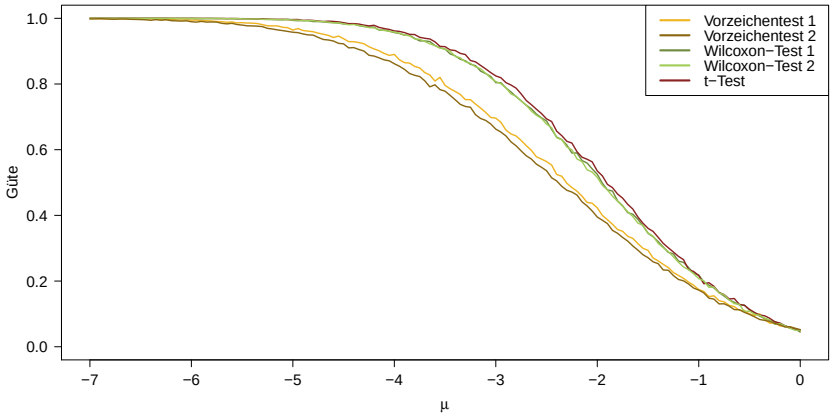

Abbildung 21: Gütefunktionen bei gerundeten Daten (ganze Zahl), S=5 und $n = 20$

3.6.5 Anwendung auf gemittelte Daten (1)

In diesem und dem nächsten Unterpunkt werden nun Daten betrachtet, die zwar ursprünglich wie gehabt aus einer Normalverteilung stammen, daraufhin aber 'unkenntlicher' gemacht wurden. So wurden die n Daten zufällig in g Gruppen geteilt und jedem der Werte der Mittelwert seiner Gruppe zugeteilt, sodass am Ende nur g unterschiedliche Werte entstehen, die jeweils $\frac{n}{g}$ Daten zugewiesen werden. Ein Blick auf die Gütefunktionen in Abbildung 22 für $n = 20$ und $g = 5$ zeigt, dass die Tests offenbar äußerst schnell erkennen, wenn sich der wahre Wert für μ von μ_0 unterscheidet. Dass dem nicht unbedingt so ist und die Höhe der Güte der einzelnen Tests hier nicht zwingend positiv anzusehen ist, zeigen die Gütewerte an der Stelle $\mu = \mu_0$; dem Punkt, bei dem eine Gütefunktion eigentlich den Fehler erster Art darstellen soll. Die Gütefunktionen liegen nämlich nicht bei $\alpha = 0.05$, sondern weit darüber um 0.2. Somit ist eine vorher festgelegte Bedingung verletzt und die Tests damit wenig brauchbar.

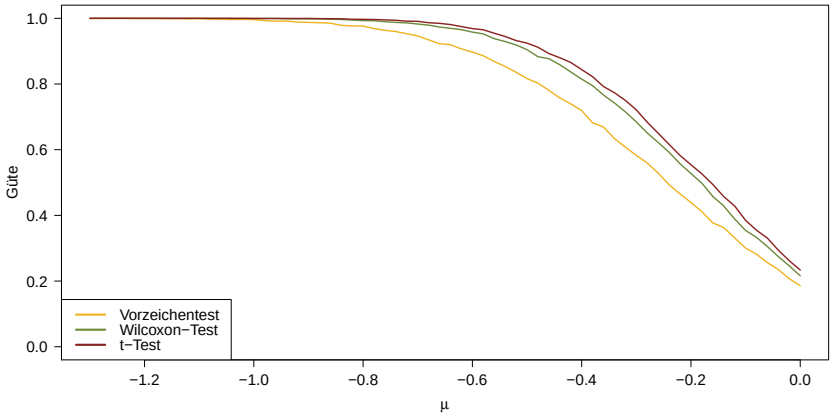

Abbildung 22: Gütefunktionen bei gemittelten Daten, $n = 20$, Gruppenanzahl =5. Die Gruppen wurden hierbei zufällig gebildet.

Warum derart hohe – 'zu hohe' – Gütewerte entstehen, lässt sich leicht erklären. Hierfür seien zufällig entstandene Werte für einen Simulationsdurchlauf bei $\mu = -0.4$ aufgeführt. Dabei muss bedacht werden, dass bei normalverteilten Daten und dem selben Stichprobenumfang die Gütefunktionen bei $\mu = -0.4$ lediglich bei etwa 0.4 (Vorzeichentest) und knapp über 0.5 (Wilcoxon- und t-Test) liegen (siehe Abbildung 3 auf Seite 14).

	Gruppe 1	Gruppe 2	Gruppe 3	Gruppe 4	Gruppe 5
Mittelwert	-0.82	-0.76	-0.60	-0.25	-0.03

Tabelle 3: Daten eines Simulationsdurchlaufs bei gemittelten Daten mit Erwartungswert $\mu = -0.4$

Entstanden sind hier nun 5 verschiedene Werte, verteilt auf jeweils 4 Daten. Der Mittelwert dieser Daten liegt hierbei bei -0.49, die Varianz knapp unter 0.1. Wie verfahren die Tests? Der Vorzeichentest erhält als Teststatistik 0 von 20 möglichen positiven Vorzeichen und lehnt ab (p-Wert: $9.5 \cdot 10^{-7}$). Für den Wilcoxon-Vorzeichen-Rang-Test ergibt sich eine positive Rangsumme von 0 und auch das führt zu einer Ablehnung der Nullhypothese (gleicher p-Wert wie bei Vorzeichentest). Der t-Test lehnt durch die kleine Varianz und einen für diese Varianz sehr stark negativen Mittelwert ab (p-Wert: $4.8 \cdot 10^{-7}$). Dies ist selbstverständlich nur ein Beispiel, jedoch zeigt es sehr realistische und gewöhnliche Daten für diese Vorgehensweise. Wie in den Gütefunktionen in Abbildung 22 zu erkennen, führt $\mu = -0.4$ bereits in den meisten Fällen zur Ablehnung der Nullhypothese, was vollkommen im Kontrast zu den Gütefunktionen der normalverteilten Daten in Abbildung 3 auf Seite 14 steht.

Neben der Nichteinhaltung des Fehler ersten Arts ist somit ein weiterer Aspekt, der gegen die Einsetzung der Hypothesentests bei vorliegenden gemittelten Daten spricht, dass hier die Gütefunktionen wesentlich höher ausfallen als bei herkömmlichen normalverteilten Daten und das alleine aus dem Grund, weil diese normalverteilten Daten (wohlgemerkt mit Informationsverlust) verfälscht worden sind.

3.6.6 Anwendung auf gemittelte Daten (2)

In diesem Abschnitt werden ursprünglich normalverteilte Daten erneut gemittelt. Der Unterschied zu dem vorangegangenen Abschnitt ist hierbei jedoch der, dass die g Gruppen der n Daten nicht zufällig gebildet werden. Stattdessen werden die Daten zunächst der Größe nach sortiert und auch so verteilt, dass die $\frac{n}{g}$ kleinsten Werte in einer Gruppe landen, die $\frac{n}{g}$ größten ebenso und so weiter.

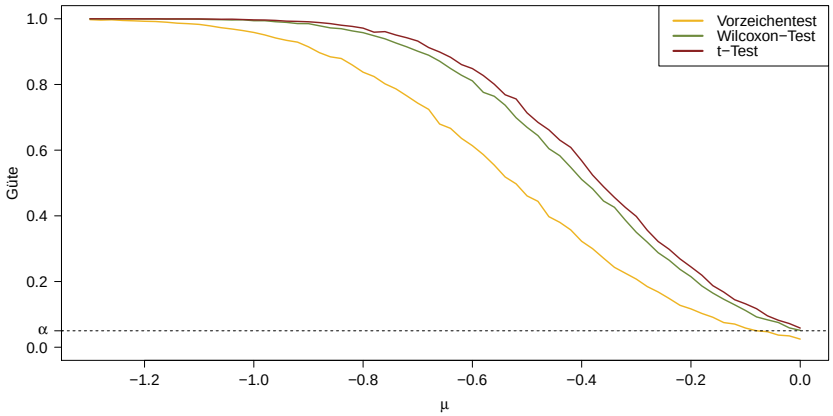

Abbildung 23: Gütefunktionen bei gemittelten Daten, $n = 20$, Gruppenanzahl =5. Die Gruppen wurden der Größe nach sortiert gebildet.

Bei einem Stichprobenumfang von 20 und 5 Gruppen sehen die Gütefunktionen ähnlich aus wie bei normalverteilten Daten, doch es ist zu erkennen, dass die Gütefunktionen des t-Tests und des Vorzeichentests an der Stelle $\mu = \mu_0$ nicht, wie eigentlich vorausgesetzt, bei $\alpha = 0.05$, sondern leicht darüber bzw. darunter liegen (Abbildung 23). Grundsätzlich verläuft die Gütefunktion des t-Tests bei diesen Daten etwas höher und die des Vorzeichentests etwas niedriger als bei normalverteilten Daten, aus denen diese gemittelten Daten entstanden sind. Der höhere Verlauf der Gütefunktion des t-Tests kann durch die entstandene niedrigere Varianz bei gleichem Mittelwert erklärt werden. Diese liegt bei diesem Stichprobenumfang und dieser Gruppenzahl bei etwa 0.9 statt 1. Dass der Vorzeichentest niedriger verläuft als bei herkömmlichen normalverteilten

Daten, ist folgendermaßen zu begründen: Das 0.05-Quantil der Teststatistik des Vorzeichentests liegt bei 6. Durch die gleichen Werte der Daten aus den jeweiligen Gruppen können nur 0, 4, 8, 12, 16 oder 20 positive Vorzeichen entstehen. 0 beispielsweise, wenn kein Gruppenmittelwert ein negatives Vorzeichen hat, was bei $\mu = \mu_0 = 0$ extrem unwahrscheinlich ist, da dafür selbst die 5 kleinsten Werte durchschnittlich größer als $\mu_0 = 0$ sein müssten. Um 4 positive Vorzeichen zu erlangen, dürfte nur ein Gruppenmittelwert größer als 0 sein, was zwar realistischer als 0 Mittelwerte, jedoch immer noch unwahrscheinlich ist. Zudem ist das α-Quantil von 6 nicht zu erreichen, da 6 kein Vielfaches von $\frac{n}{q} = 4$ ist und so keine Randomisierung stattfinden kann, die die Wahrscheinlichkeit für die Ablehnung der Nullhypothese erhöht hätte.

Unterschiede bestehen, wenn als Stichprobenumfang 10 und als Gruppenanzahl wieder 5 gewählt wird, wie in Abbildung 24 zu betrachten ist. Hier ist die Güte des Vorzeichentests an der Stelle $\mu = \mu_0$ größer als α. In diesem Fall lässt sich das 0.05-Quantil von 2 durchaus erreichen und durch die Randomisierung wird die Nullhypothese dadurch mit einer Wahrscheinlichkeit von 89% abgelehnt (siehe Abschnitt 2.4). Wieder zu sehen ist, dass der Fehler erster Art des t-Tests bei $\mu = \mu_0$ leicht über α liegt, da die Varianz mit 0.92 wieder leicht kleiner ist als 1.

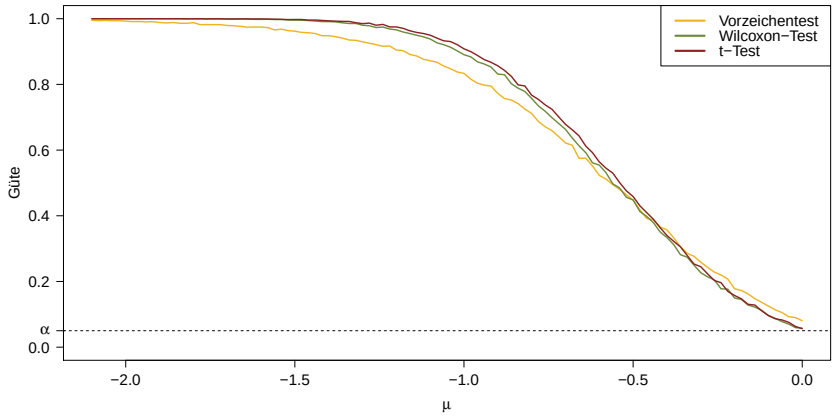

Abbildung 24: Gütefunktionen bei gemittelten Daten, $n = 10$, Gruppenanzahl = 5. Die Gruppen wurden der Größe nach sortiert gebildet.

Grundsätzlich ist zu den der Größe nach sortiert gemittelten Daten zu sagen, dass sie die Teststatistiken der nonparametrischen Tests und deren Verteilungen verfälschen, indem viele potentielle Ausprägungen hiervon nicht mehr realisierbar und auch die p-Werte der noch möglichen Ausprägungen nicht mehr die selben sind. Bei dem Wilcoxon-Vorzeichen-Rang-Test wirkt sich dieses Problem

bereits bei einer recht kleinen Gruppenanzahl g nur kaum sichtbar aus. Immerhin sind dennoch 2^g verschiedene Rangsummen möglich, was eine Erklärung für die nur minimale Verfälschung sein könnte.

In Abbildung 25 beträgt der Stichprobenumfang 15 und die Gruppenanzahl lediglich 3. Hier zeigt sich, dass auch beim Wilcoxon-Test nun an der Stelle $\mu = \mu_0$ weniger als 5% der Durchführungen zu einer Ablehnung der Nullhypothese führen.

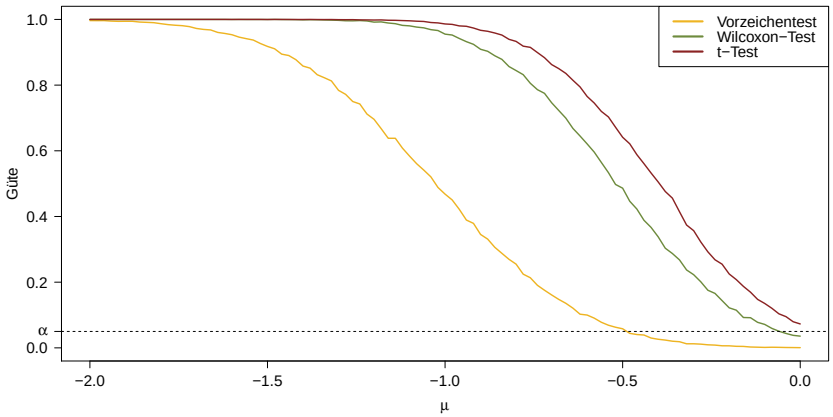

Abbildung 25: Gütefunktionen bei gemittelten Daten, $n = 15$, Gruppenanzahl = 3. Die Gruppen wurden der Größe nach sortiert gebildet.

Zusätzlich ist zu sehen, dass die Güte des t-Tests bei einer Varianz von 0.8 erneut an der Stelle $\mu = \mu_0$ über 0.05 liegt. Die Gütefunktion des Vorzeichentests ist lange unter 0.05 und geht von links bis zum Punkt 0 auf minimal über 0 zu. Dies liegt daran, dass für die Ablehnung der Nullhypothese in diesem Fall bei einem 5%-Quantil von 4 und möglichen Ausprägungen der Teststatistik von 0, 5, 10 und 15 kein Gruppenmittelwert über 0 sein darf, was wiederum extrem unwahrscheinlich ist.

Zusammengefasst ist also zu sagen, dass der t-Test in diesem Fall weniger konservativ ist, als er eigentlich sein sollte. Dies liegt wie erwähnt an der kleineren Varianz trotz des identischen Mittelwerts. Die Verteilungen der Teststatistiken der nonparametrischen Tests werden deutlich verändert und die möglichen Ausprägungen der Teststatistik reduziert, was sich vor allem bei dem Vorzeichentest auswirkt.

4 Simulation der finiten relativen Effizienz

Im vorangegangenen Kapitel wurden zahlreiche Gütefunktionen dargestellt, die zu konkreten Datensituationen einen relativ genauen Einblick über die Power von Tests liefern. Ein Nachteil dieser Gütefunktionen ist es allerdings, dass dafür eine genaue, intensivere Betrachtung notwendig ist und dabei meist nicht eindeutig auf den Punkt gebracht werden kann, in welchen Situationen sich die Tests wie stark voneinander unterscheiden. Nur bei großen Differenzen können auf den ersten Blick Aussagen getroffen werden wie, dass ein Test in einer Situation noch stärker überlegen ist als in einer anderen. In den Abbildungen 6 (Seite 17) und 9 (Seite 20) ist beispielsweise die stärkere Unterlegenheit des t-Tests bei trimodalen ($\epsilon = 10\%$) im Vergleich zu laplaceverteilten Daten schnell und ohne große Zweifel erkennbar, doch das stellt eher eine Ausnahme dar.

Aus diesen Gründen ist eine Kennzahl wünschenswert, die die Stärke der Güteunterschiede angeben kann. Eine Möglichkeit dafür ist die finite relative Effizienz, die in diesem Kapitel zunächst definiert und daraufhin simuliert wird.

4.1 Definition und Verwirklichung

Kurz zusammengefasst ist die finite relative Effizienz das Verhältnis zweier Stichprobenumfänge, die bei den zu vergleichenden Tests jeweils nötig sind, um eine gewisse Güte zu erreichen (Büning & Trenkler; 1994, S. 276 ff.). Festgelegt werden muss dabei das Signifikanzniveau α, der wahre Parameter μ und die angestrebte Güte $G(\mu)$. Dafür werden erneut Gütefunktionen erstellt, hier allerdings in Abhängigkeit des Stichprobenumfangs n. Für α wird wieder ein Wert von 5% gewählt. Die Güte, die erreicht werden soll, wird auf 0.8 gesetzt, da nach Jacob Cohen der Fehler zweiter Art maximal vier mal so groß sein soll wie der Fehler erster Art (Cohen; 1988, S. 390 ff.). Für μ wird kein gleichbleibender Wert festgelegt, da die Daten in den verschiedenen Situationen unterschiedlich stark streuen und ein über alle Situationen gleichbleibender Abstand $\mu - \mu_0$ unterschiedlich groß erscheint. Um möglichst einheitlich zu bleiben, wurde sich dafür entschieden, $-\mu$ auf die halbe erwartete Standardabweichung der jeweiligen Daten zu setzen.

In der Simulation werden nun die Werte der Gütefunktionen betrachtet und für jeden Test der kleinste Stichprobenumfang ausgemacht, der unter den anderen festen Größen zu einer Güte von mindestens 0.8 führt. Daraufhin werden diese verschiedenen Stichprobenumfänge in Relation gesetzt. Abbildung 26 zeigt ein Beispiel hierfür an normalverteilten Daten mit $\mu = -0.5$.

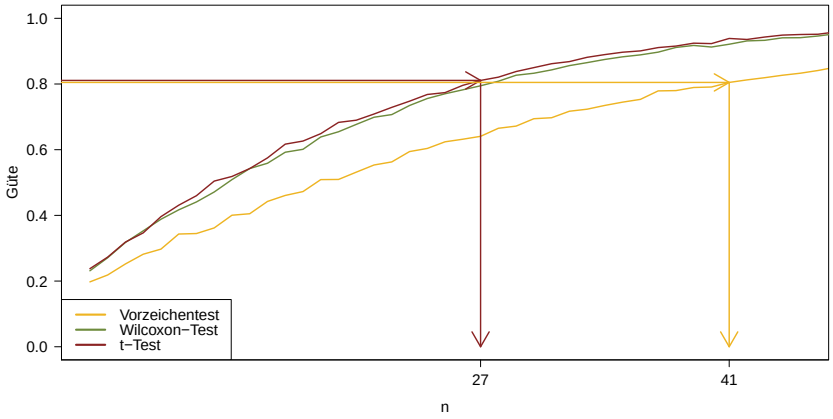

Abbildung 26: Gütefunktionen bei normalverteilten Daten in Abhängigkeit von n. Der kleinste Stichprobenumfang, der bei dem t-Test zu einer Güte von über 0.8 führt, ist 27, beim Vorzeichentest 41. Die simulierte, finite relative Effizienz des Vorzeichentests beträgt in diesem Fall $\frac{27}{41} \approx 0.66$

Die gefragten Stichprobenumfänge zweier Tests werden im Folgenden $n_1(G)$ und $n_2(G)$ genannt und sind also folgendermaßen definiert:

$$n_1(G) = min\{n' \in \mathbb{N} | G_1(n', \alpha = 0.05, \mu) \geq 0.8\}$$
$$n_2(G) = min\{n'' \in \mathbb{N} | G_2(n'', \alpha = 0.05, \mu) \geq 0.8\}$$

Der Quotient $\frac{n_1(G)}{n_2(G)}$ wird dann als relative Effizienz des ersten im Verhältnis zum zweiten Test bezeichnet (Büning & Trenkler; 1994, S. 278). Liegt dieser Wert über 1 und ist damit der benötigte Stichprobenumfang für den ersten Test größer, so wird dieser Test als weniger effizient bezeichnet. Da es in Simulationen trotz der hohen Anzahl an Durchläufen vorkommen kann, dass auf den ersten Gütewert über 0.8 wieder ein Wert von unter 0.8 folgt, wurde sich hier dafür entschieden, dass in diesem Fall $n_i(G)$ um 1 erhöht wird. Die Wahrscheinlichkeit für das Eintreffen dieses Problems ist jedoch sehr gering.

Zu den möglichen Realisationen für die simulierte finite relative Effizienz sei gesagt, dass sie als das Verhältnis zweier natürlicher Zahlen nur begrenzt viele Werte annehmen kann. Problematisch wird dies vor allem dann, wenn zum Erreichen der gefragten Güte jeweils nur sehr kleine Stichprobenumfänge benötigt werden. Bei äußerst kleinen Güteunterschieden, die in der Grafik jedoch noch beobachtbar sind, ergibt sich so eventuell eine relative Effizienz von 1. Da – wie sich zeigen wird – für die halbe mittlere Standardabweichung als wahres μ ein relativ hoher Stichprobenumfang zur Erreichung einer Güte von 80% benötigt wird, erscheint diese Wahl aber als durchaus passend. Die Wahl

eines betragsmäßig noch kleineren μ gewährte zwar eine feinere Abstufung zwischen den potentiellen Werten, in einigen Fällen wurden Effizienzunterschiede allerdings eindeutig unterschätzt, weshalb diese Vorgehensweise nicht zur Anwendung kommt. Zudem ist der Vergleich zu den Grafiken einfacher, da bei einem Stichprobenumfang von 20 die Abweichung zu μ_0 von der halben Standardabweichung zumeist etwa in der Mitte der Grafik angesiedelt ist.

In den nachfolgenden Simulationen wurde zwischen den beiden verschiedenen Varianten des Wilcoxon-Tests nur unterschieden, falls Nulldifferenzen vorliegen. Als Begründung hierfür kann wieder auf Abschnitt 3.3 verwiesen werden. Für die Datenverteilungen, deren Gütefunktionen bereits als problematisch besprochen wurden, wurde die relative finite Effizienz nicht simuliert.

4.2 Ergebnisse

Nun sollen die Ergebnisse dieser Simulationen vorgestellt werden. Für die Anzahl der Simulationsdurchläufe fiel die Entscheidung wie gehabt auf den Wert 10000. Bei kleineren Werten war die Wahrscheinlichkeit für zufällige Abweichungen zu hoch und damit entstanden in nicht wenigen Ausnahmefällen unrealistische Effizienzwerte. Die berechnete Effizienz ist in der folgenden Tabelle immer als Verhältnis des benötigten Stichprobenumfangs eines Tests dividiert durch jenen des t-Tests zu verstehen. Sind also Werte über 1 zu finden, so ist der t-Test dem entsprechenden Test im Vergleich unterlegen.

Die Abkürzungen in der nachfolgenden Tabelle 4 sind folgendermaßen zu verstehen:

- 'V1' und 'V2' stehen für die Varianten des Vorzeichentests. Die Nummerierung ist wie bei den Grafiken zu verstehen. Gleiches gilt für die Abkürzungen für den Wilcoxon-Test ('W1' und 'W2').
- N $\widehat{=}$ Normalverteilte Daten (siehe 3.4)
- Unif $\widehat{=}$ Stetig gleichverteilte Daten (siehe 3.5.1)
- L $\widehat{=}$ Laplaceverteilte Daten (siehe 3.5.2)
- N/L $\widehat{=}$ Mischung aus normal- und laplaceverteilten Daten (siehe 3.5.3)
- Unif/L $\widehat{=}$ Mischung aus stetig gleich- und laplaceverteilten Daten (siehe 3.5.3)
- 1 NKS $\widehat{=}$ Auf eine Nachkommastelle gerundete Daten (siehe 3.6.4
- Ganze Zahl $\widehat{=}$ Auf die ganze Zahl gerundete Daten (siehe 3.6.4)
- Trimodal $\widehat{=}$ Trimodale Daten mit entsprechend angegebenen Werten für ϵ und ξ (siehe 3.6.1)

	V1	V2	W1	W2
N	0.66	/	0.96	/
Unif	0.36	/	0.81	/
L	1.13	/	1.18	/
N/L	0.87	/	1.04	/
Unif/L	0.70	/	0.96	/
1 NKS ($\sqrt{Var} = 1$)	0.69	0.66	0.96	0.96
Ganze Zahl ($\sqrt{Var} = 1$)	0.88	0.55	0.91	0.97
Ganze Zahl ($\sqrt{Var} = 3$)	0.75	0.64	0.96	1
Ganze Zahl ($\sqrt{Var} = 5$)	0.71	0.66	0.96	1
Trimodal ($\epsilon = 0.1$, $\xi = 10$)	3.38	/	2.70	/
Trimodal ($\epsilon = 0.25$, $\xi = 10$)	2.70	/	1.29	/
Trimodal ($\epsilon = 0.1$, $\xi = 3$)	0.96	/	1.17	/
Trimodal ($\epsilon = 0.25$, $\xi = 3$)	1.08	/	1.08	/
Trimodal ($\epsilon = 0.5$, $\xi = 3$)	0.73	/	0.90	/
Trimodal ($\epsilon = 0.75$, $\xi = 3$)	0.28	/	0.87	/

Tabelle 4: Simulierte finite relative Effizienz einiger Datensituationen. μ ist hierbei jeweils die halbe mittlere Standardabweichung der jeweiligen Daten.

Zunächst ist zu sagen, dass sich die simulierten Werte für die finite relative Effizienz grundsätzlich gut mit den Gütefunktionen decken. Die simulierten Werte sind plausibel und Güteunterschiede zwischen den einzelnen Tests und auch den einzelnen Datensituationen können damit gut, kurz und knapp dargestellt werden. Beispielsweise sind sehr niedrige Werte bei der finiten relativen Effizienz bei stetig gleichverteilten Daten vorzufinden, so wie es sich bereits grafisch gezeigt hat. Trotz der Vorzüge der simulierten Kennzahlen ergeben sich hierbei jedoch noch Probleme:

1. Als Effizienz ergibt sich teilweise 1, obwohl die dazugehörigen Plots der Gütefunktionen eindeutig die Verschiedenheit von 1 zeigen. Der Unterschied zwischen den Gütewerten ist nur so gering, dass der minimale Stichprobenumfang zum Erreichen der geforderten Güte von zwei Tests genau gleich ist. Dass dieses Problem entstehen kann, wurde bereits im vorangegangenen Punkt 4.1 erläutert.

2. Der Wert 1.29 der simulierten finiten relativen Effizienz des Wilcoxon-Tests bei trimodalen Daten mit $\epsilon = 0.25$ und $\xi = 10$ ist mit erhöhter Vorsicht zu genießen (siehe Abbildung 10 auf Seite 22). Das für die simulierte Effizienz gewählte $\mu = -\frac{\sqrt{11}}{2} \approx 1.66$ befindet sich in dem Bereich, in dem die Güte des Wilcoxon-Tests stagniert. Ist μ beispielsweise als -4 festgelegt, so wäre eine Effizienz von unter 1 zu erwarten, da in diesem Bereich der t-Test dem Wilcoxon-Test überlegen ist. Aus diesem Grund sind etwa bei den tri- und bimodalen Daten die simulierten Werte der fini-

ten relativen Effizienz nicht für alle möglichen Kombinationsmöglichkeiten von ϵ und ξ aufgeführt.

3. Bei mehreren Durchläufen kann es durchaus vorkommen, dass sich andere Werte ergeben. Diese sind den hier aufgeführten Werten jedoch relativ ähnlich und so sind die sich ergebenden Zahlen als grober Richtwert zu verstehen.

5 Fazit und Ausblick

Sicher ist eines: Da in den vorliegenden Simulationen immer einige Parameter mehr oder minder vollkommen willkürlich festgesetzt werden können bzw. müssen, besteht unendlich viel Spielraum und allgemeine Schlüsse können sozusagen überhaupt nicht gezogen werden. Gerade im Umfang einer solchen Arbeit ist es nicht möglich, einige Parameter mehr als nur ein wenig zu variieren, was womöglich dazu führt, dass vielleicht wichtige Erkenntnisse übersehen werden. Außerdem bleibt immer zu sagen, dass Simulationen Daten auf eine zuvor 'festgelegte' und damit stark berechenbare Art und Weise generieren. In der Realität folgen Daten jedoch höchstens annähernd einer theoretischen Verteilung und es ist genau das Problem, dass die wahre Verteilung nicht eindeutig bestimm- oder berechenbar verläuft. Aus diesen Gründen können durch Simulationen nur in einem begrenzten Rahmen Schlüsse gezogen werden, die wiederum nicht zwingend allgemeingültig sind. Daher bleibt zu sagen, dass sie nur einen kleinen theoretischen Einblick über die Güte von Tests liefern können. Auch eine Kennzahl über die Güte zu simulieren, erweist sich anhand der finiten relativen Effizienz als nicht unproblematisch. Andere allgemeingültigere Methodiken hierzu existieren zwar auch (siehe Büning & Trenkler (1994, S. 279 ff.)), dort ergeben sich jedoch andere Probleme wie mangelnde Genauigkeit und die äußerst schwere Berechnung. Trotz der angesprochenen Problematiken konnten dennoch eindeutige Ergebnisse geliefert werden, nach denen sich mindestens grob gerichtet werden kann.

Im Rahmen dieser Arbeit zeigt sich so, dass der t-Test beiden nonparametrischen Tests in der Mehrheit der Fälle überlegen ist. Dabei sei seine häufige Überlegenheit vor allem unter dem Gesichtspunkt bedacht, dass seine Voraussetzung der Normalverteilung der Daten meist verletzt wurde. Eine stetige Gleichverteilung beispielsweise gleicht der Normalverteilung kaum, jedoch ist die Güte des t-Tests hier der der anderen Tests mit am stärksten überlegen. In den allermeisten betrachteten Fällen ergab sich, dass die nonparametrischen Tests entweder beide besser oder beide schlechter abschnitten als der t-Test, was vermutlich der ähnlichen Methodik des Vorzeichen- und des Wilcoxon-Vorzeichen-Rang-Tests zuzuschreiben ist. Wilcoxons Test hat in den meisten betrachteten Fällen eine höhere Güte als der Vorzeichentest und büßt im Vergleich zum t-Test – wenn überhaupt – zumeist überschaubar viel Güte ein. Dies belegen auch die simulierten Werte für die finite relative Effizienz, die sich mit den Ergebnissen aus den Gütekurven trotz vorhandener Problematiken gut decken.

Was den Informationsverlust durch die Vorzeichen- und Rangbildung betrifft, lässt sich Folgendes sagen: Grundsätzlich können die nonparametrischen Tests robuster mit Ausreißern bzw. extremen Werten umgehen. Es zeigt sich jedoch, dass beispielsweise durch solche die Gütefunktionen des Wilcoxon-Tests und des Vorzeichentests auch systematisch beeinflusst werden können und diese nicht mehr zwischen verschiedenen wahren Erwartungswerten unterscheiden können.

Zusammengefasst resultiert aus dieser Arbeit, dass der Wilcoxon-Vorzeichen-Rang-Test trotz oder gerade wegen des durch seine Methodik bedingten Informa-

tionsverlusts der Daten mindestens eine relativ sichere – und im Vergleich zum Vorzeichentest meist bessere – Alternative zum t-Test bei Unsicherheit über die Verteilung einer Stichprobe ist. Hin und wieder ist er dem t-Test überlegen und wenn nicht, hält sich der damit verbundene Effizienzverlust noch im Rahmen.

A Weitere Abbildungen

A.1 Weitere Gütefunktionen zu Daten anderer Verteilungen

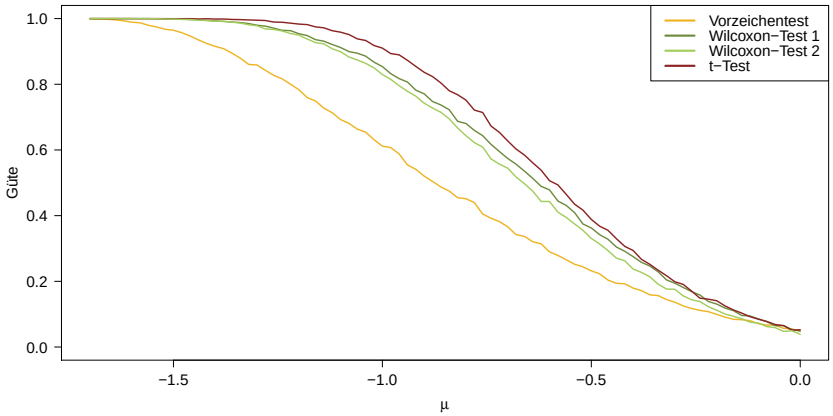

Abbildung 27: Gütefunktionen bei stetig gleichverteilten Daten und $n = 10$

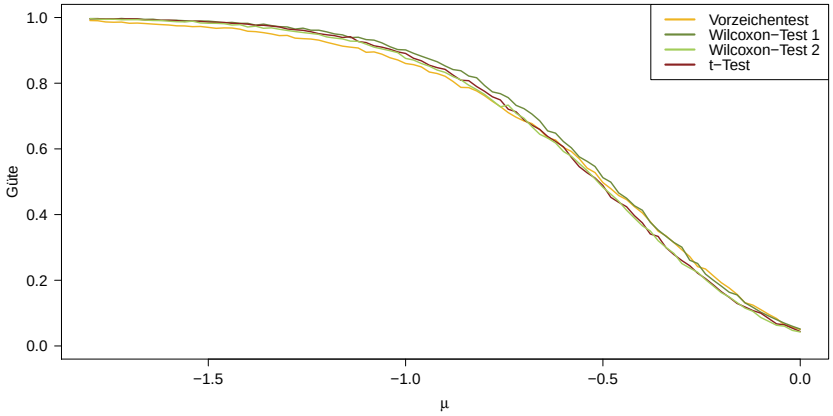

Abbildung 28: Gütefunktionen bei laplaceverteilten Daten und $n = 10$

A.2 Weitere Gütefunktionen zu tri- und bimodalen Daten

Im Abschnitt 3.6.1 mit tri- und bimodalen Daten wurde für ξ der sehr große Wert 10 gewählt. Im Folgenden sind die Gütefunktionen mit entsprechenden ϵ

zu finden, bei denen allerdings $\xi = 3$ gilt. Die Unterschiede bei verschiedenen Werten für ξ zu den Gütefunktionen sind, wie unschwer zu erkennen, sehr groß.

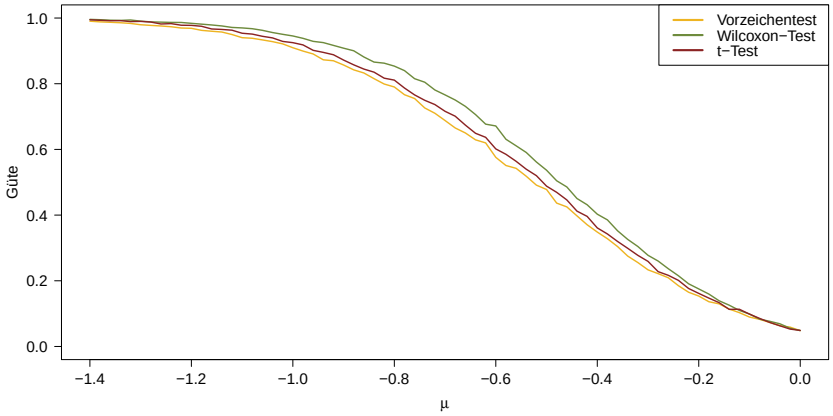

Abbildung 29: Gütefunktionen bei trimodalen Daten ($\xi = 3$, $\epsilon = 10\%$, $n = 20$). Auffällig ist hier, dass der t-Test bei weitem nicht so stark abfällt wie bei $\xi = 10$ und seine Gütefunktion sogar knapp über der des Vorzeichentests liegt.

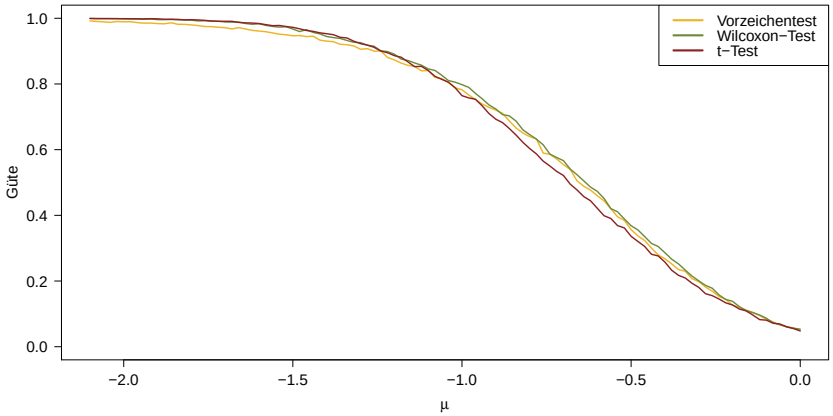

Abbildung 30: Gütefunktionen bei trimodalen Daten ($\xi = 3$, $\epsilon = 25\%$, $n = 20$). Während bei $\xi = 10$ bereits ein stellenweiser konstanter Verlauf der Gütefunktion des Wilcoxon-Vorzeichen-Rang-Tests zu sehen ist, ist dies hier nicht zu beobachten. Hier ist die Güte des Vorzeichentests jedoch wieder höher als die des t-Tests.

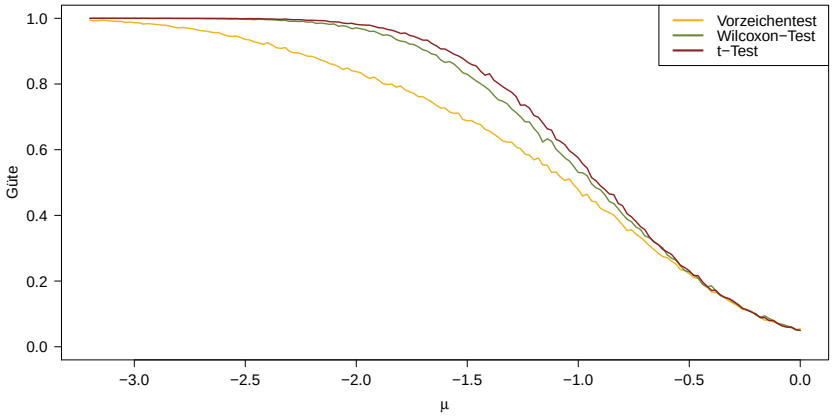

Abbildung 31: Gütefunktionen bei trimodalen Daten ($\xi = 3$, $\epsilon = 50\%$, $n = 20$). Die Gütefunktionen verlaufen nach wie vor in einer herkömmlichen Form und der t-Test ist den nonparametrischen Tests sogar überlegen.

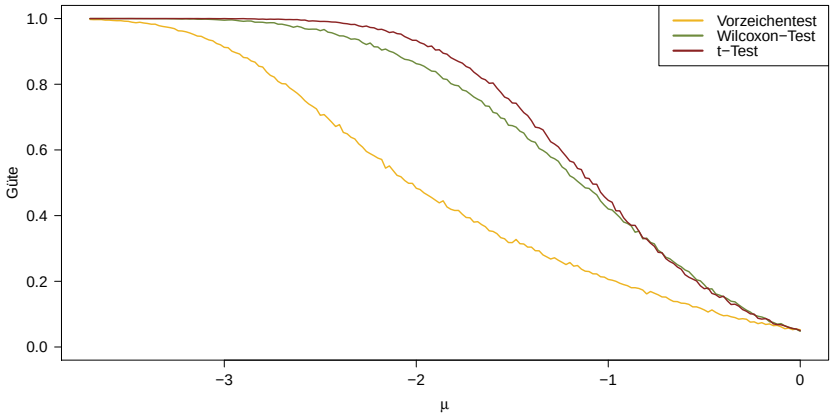

Abbildung 32: Gütefunktionen bei trimodalen Daten ($\xi = 3$, $\epsilon = 75\%$, $n = 20$). Nach wie vor sind keine stellenweise konstant verlaufenden Gütefunktionen zu sehen. Der Vorzeichentest kann in seiner Güte absolut nicht mehr mit den anderen Tests mithalten.

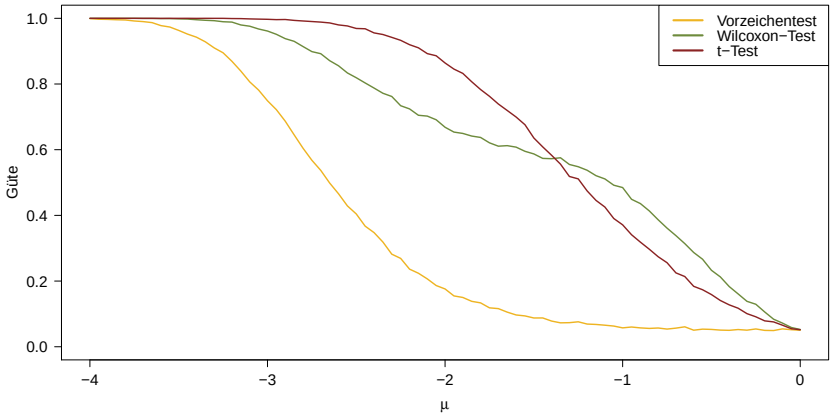

Abbildung 33: Gütefunktionen bei bimodalen Daten ($\xi = 3$, $\epsilon = 100\%$, $n = 20$). Hier ist nun ein stellenweise konstanter Verlauf beim Wilcoxon-Test etwa zwischen $\mu = -2$ und $\mu = -1$ zu erahnen. Seine und die Gütefunktion des t-Tests überschneiden sich in diesem Bereich. Der Vorzeichentest kann bis zu einem sehr klein gewählten μ das Zutreffen der Alternativhypothese nicht erkennen und seine Gütefunktion stagniert auf dem Niveau α

A.3 Weitere Abbildungen zu gerundeten Daten

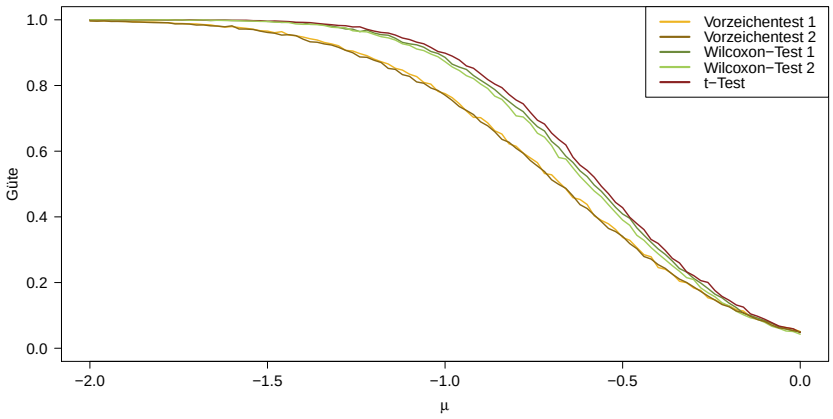

Abbildung 34: Gütefunktionen bei gerundeten Daten (eine Nachkommastelle) und $n = 10$

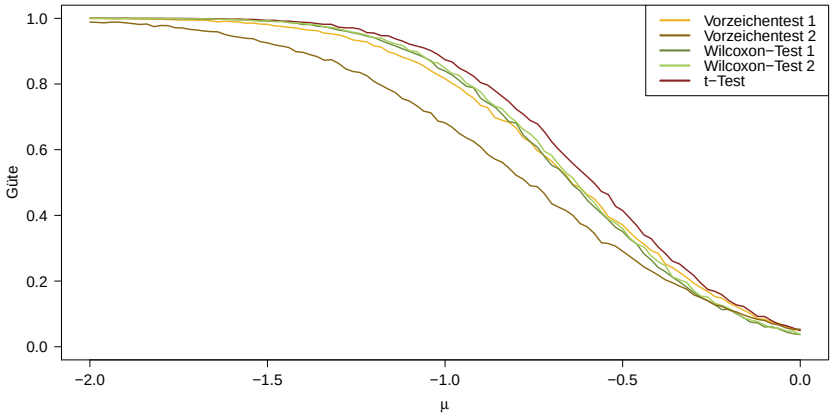

Abbildung 35: Gütefunktionen bei gerundeten Daten (ganze Zahl) und $n = 10$

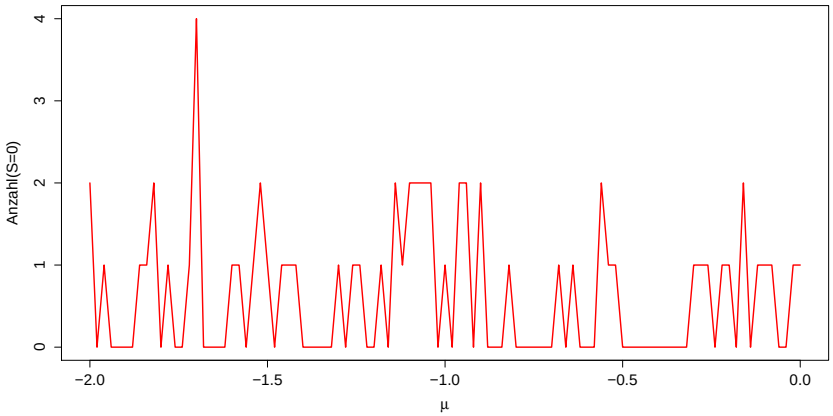

Abbildung 36: Anteil der 10000 Simulationsdurchläufe bei auf ganze Zahlen gerundete Daten mit Stichprobenumfang $n = 10$ (Gütefunktion siehe Abbildung 35), in denen sich $S = 0$ ergibt. Berechnet werden kann die Teststatistik des t-Tests nur, wenn S in keinem der Durchläufe 0 ist, was in dem betrachteten Fall jedoch sehr häufig nicht der Fall ist. Außerdem ist allerdings zu sehen, dass die Anzahl der Fälle, bei denen sich für die Stichprobenvarianz der Wert 0 ergibt, gemessen an 10000 Simulationsdurchläufen, bei jeder Wahl von μ sehr gering ist.

A.4 Gütefunktionen mit wenigen Simulationsdurchläufen

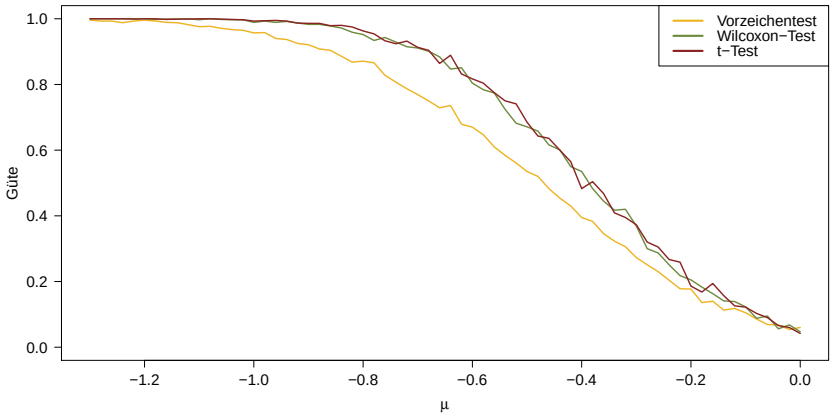

Abbildung 37: Gütefunktionen bei normalverteilten Daten ($n = 20$). Hier wurden nur 1000 Simulationsdurchläufe durchgeführt, was sich an den absolut nicht glatten Gütekurven bemerkbar macht. Während bei 10000 Durchläufen klar ausgemacht werden kann, dass der t-Test eine höhere Güte als der Wilcoxon-Test besitzt (siehe Abbildung 3), überschneiden sich deren Kurven hier oftmals durch zufällige Abweichungen und die Überlegenheit des t-Tests ist grafisch nicht mehr wirklich bestimmbar.

Literatur

Abramowitz, M. & Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth gpo printing edn, Dover, New York.

Büning, H. & Trenkler, G. (1994). *Nichtparametrische statistische Methoden*, 2. edn, de Gruyter.

Bortz, J., Lienert, G. & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik*, Springer-Lehrbuch, 3. edn, Springer.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2 edn, L. Erlbaum Associates.

Duller, C. (2008). *Einführung in die nichtparametrische Statistik mit SAS und R: Ein anwendungsorientiertes Lehr- und Arbeitsbuch*, Physica-Lehrbuch, Physica-Verlag HD.

Fahrmeir, L., Pigeot, I. & Tutz, G. (2007). *Statistik. Der Weg zur Datenanalyse*, 6. edn, Springer Verlag, München.

Hennig, C. (2012). *R-Package 'smoothmest': Smoothed M-estimators for 1-dimensional location.*

Kauermann, G. & Hothorn, T. (2014). *Statistik IV Modul P8: Grundlagen der Statistik II Vorlesung P8.1: Wahrscheinlichkeitstheorie und Inferenz II.*
URL: *http://www.statistik.lmu.de/institut/lehrstuhl/wisoz/lehre/stat4_ss14/download/Statistik_IV.pdf*

Pagenkopf, J. (1977). *Güte und Effizienz einiger nicht-parametrischer Tests bei kleinen Stichproben*, number 5 in *Studien zur angewandten Wirtschaftsforschung und Statistik aus dem Institut für Statistik und Ökonometrie der Universität Hamburg*, 1. edn, Vandenhoeck & Ruprecht, Göttingen.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Wilcoxon, F. (1945). Individual comparisons by ranking methods, *Biometrics Bulletin* **1**(6): 80–83.

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Erklärung

Ich versichere hiermit, dass ich die vorliegende Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 11. März 2015

Tobias Steinherr