Technische Universität Dortmund

Fakultät für Statistik

Lehrstuhl für Statistik mit Anwendungen im Bereich der Ingenieurwissenschaften

Bachelorarbeit

Vergleich von Zwei-Stichproben-Relevanz-Tests basierend auf t-Tests und Datentiefen

verfasst von Dennis Andreas Malcherczyk

Betreuerin:
Prof. Dr. Christine Müller

27. Juni 2018

Inhaltsverzeichnis

1 Einleitung

Der Vergleich von Stichproben aus verschiedenen Gruppen taucht in den unterschiedlichsten Problemstellungen der statistischen Anwendung auf. Zum Beispiel werden in klinischen Studien mehrere Versuchsgruppen mit unterschiedlichen Therapien miteinander verglichen, in technischen Anwendungen wird der Verschleiß von Bauteilen in der Entwicklung mehrerer Zeitperioden untersucht oder man ist am Vergleich von verschiedenen Kursentwicklungen in der Wirtschaft interessiert.
In dieser Arbeit werden Testverfahren für zwei unabhängige Stichproben vorgestellt. Diese Testverfahren liefern anhand von erhobenen Daten Entscheidungsregeln für die Fragestellung, ob die Mittelwerte beider Stichproben einen signifikanten Unterschied besitzen. Zudem möchte man keine zu kleinen Unterschiede zulassen, sondern nur relevante Unterschiede feststellen. Die Größe dieses relevanten Unterschieds kann vom Anwender frei ausgewählt werden, um je nach Sachkontext keine zu kleinen Unterschiede als signifikant zu bezeichnen. Dieser Gedanke führt zu sogenannten *Relevanz-Tests* für zwei Stichproben, die den Kerngegenstand dieser Arbeit bilden.
Im Abschnitt 2 werden verschiedene parametrische und nichtparametrische Relevanz-Tests vorgestellt. Die parametrischen Relevanz-Tests entsprechen Verallgemeinerungen des klassischen Zweistichproben-t-Tests. Die nichtparametrischen Relevanz-Tests beruhen auf Datentiefen, Maßzahlen mit robusten Eigenschaften, die aus den Vorzeichen der Abweichungen vom Mittelwert berechnet werden. Dabei lässt sich eine der beiden verwendeten Vorzeichen-Tiefen in vielen Modellen aus der Simplex-Regressions-Tiefe von Rousseeuw und Hubert (1999) und Müller (2005) ableiten, siehe Kustosz, Müller, Wendler (2016). Die andere Vorzeichen-Tiefe entspricht einer vereinfachten Version der aus der Simplex-Regressions-Tiefe abgeleiteten Vorzeichen-Tiefe.
Die Einführung verschiedener Testverfahren wirft die Frage auf, welcher der Tests verwendet werden sollte. Daher werden im Abschnitt 3 die Relevanz-Tests in einer Simulationsstudie unter verschiedenen Modellannahmen miteinander verglichen. Abhängig vom Modell können sowohl die parametrischen als auch die nichtparametrischen Tests bessere Ergebnisse erzielen. Im letzten Abschnitt 4 werden die wich-

tigsten Ergebnisse zusammengefasst und diskutiert. Insbesondere folgt ein Ausblick für weitere Untersuchungsmöglichkeiten zu noch offen stehenden Fragen.

Zum Abschluss der Einleitung soll besonders Prof. Dr. Christine Müller gedankt werden, welche die Erstellung dieser Arbeit erst ermöglichte und sie mit Anregungen, Beantwortung von Fragen aller Art und Kritik in einem sehr bereichernden Umfang betreute.

2 Statistische Verfahren und Grundlagen

Zunächst sollen die Generalannahmen und Notationen dieser Arbeit vorgestellt werden. Seien sowohl $X_1, ..., X_M$ als auch $Y_1, ..., Y_N$ unabhängig, identisch verteilte Zufallsvariablen. Sie stellen die beiden Stichproben dar. Gelegentlich wird für den Vektor der Stichproben auch $X = (X_1, ..., X_M)$ und $Y = (Y_1, ..., Y_N)$ geschrieben. Wenn nichts anderes gesagt wird, sind M und N die Größen der Stichproben, wobei oft $M = N$ angenommen wird. Die beiden Stichproben X, Y seien stets zueinander stochastisch unabhängig. Die Zufallsvariablen sollen folgende Gestalt besitzen:

$$X_i = \mu_1 + E_i \quad \text{für alle } i = 1, ..., M, \tag{1}$$

$$Y_j = \mu_2 + E_{M+j} \quad \text{für alle } j = 1, ..., N. \tag{2}$$

Dabei seien $E_1, ..., E_{M+N}$ unabhängig, identisch verteilte Zufallsvariablen mit

$$\text{med}(E_i) = 0 \quad \text{für alle } i = 1, ..., M + N.$$

Bei diesem Modellierungsansatz wird von einem deterministischen Mittelwert μ_1 bzw. μ_2 ausgegangen, auf dem die Zufallsvariablen $E_1, ..., E_{M+N}$ als stochastische Fehler einwirken. Für beide Stichproben wird angenommen, dass die stochastischen Abweichungen sich identisch verhalten. Durch diesen Modellierungsansatz können sich die Verteilungen der Stichproben höchstens um ihren Lageparameter unterscheiden. Die Forderung an den Median für E_i mit $i = 1, ..., M + N$ unterbindet, dass zusätzlich systematische Lageverschiebungen durch die stochastischen Fehler hervorgerufen werden. Außerdem kann man so erwarten, dass im Mittel genauso viele Realisationen über und unter dem Lageparameter sind. Diese Bedingung wird später bei der Konstruktion der Relevanz-Tests mit der Datentiefe eine Rolle spielen.
Zweistichproben-Tests befassen sich mit der Frage, ob eine signifikante Abweichung zwischen den unbekannten Mittelwerten μ_1 und μ_2 vorliegt. Durch folgendes Hypothesenpaar lässt sich die Fragestellung wie folgt formulieren:

$$H_0 : \mu_1 = \mu_2 \quad \text{gegen} \quad H_1 : \mu_1 \neq \mu_2. \tag{3}$$

In dieser Arbeit wird diese Problemstellung in einem allgemeineren Rahmen betrachtet. Dabei lässt man kleine Unterschiede von geringer Relevanz zwischen den Mittelwerten zu. Der relevante Unterschied wird durch einen Parameter $\delta \geq 0$ festgelegt. Anschließend lässt sich das Hypothesenpaar verallgemeinert darstellen:

$$H_0 : |\mu_1 - \mu_2| \leq \delta \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > \delta. \tag{H}$$

Bei bestimmten Sachzusammenhängen ist es sinnvoll sehr kleine Abweichungen zuzulassen. Vor allem bei hohen Stichprobenumfängen können im Hypothesenpaar in Formel (3) immer kleinere, aber eigentlich irrelevante Abweichungen zunehmend besser erkannt werden. Um solche irrelevante Effekte zu kontrollieren, werden Relevanz-Tests eingesetzt (Müller und Denecke (2013), S. 223). Alle Testverfahren dieser Arbeit untersuchen im Wesentlichen das Hypothesenpaar (H). Dabei wird der Annahmebereich des Tests zu (H) als *Nicht-Relevanzbereich* bezeichnet.
Die im Kapitel 2 vorgestellten Testverfahren werden schrittweise aufgebaut. Es werden verschiedene Zweistichproben-Relevanz-Tests unter Normalverteilungsannahme vorgeschlagen. In Abschnitt 2.1 werden die Standard-t-Tests für unabhängige Stichproben thematisiert. In Abschnitt 2.2 werden die t-Tests mit Berücksichtigung von relevanten Unterschiedenen aus den standardmäßigen t-Tests hergeleitet.
Nichtparametrische Relevanz-Tests unter Verwendung von Datentiefen werden in Abschnitt 2.3 vorgestellt. Testverfahren mittels Datentiefen sind einerseits robust gegenüber extremen Werten (insbesondere Ausreißern). Andererseits müssen keine expliziten Verteilungsannahmen der E_i für $i = 1, ..., M+N$ wegen der Nichtparametrik getroffen werden. Daher liefern neben normalverteilten Fehlervariablen auch andere Verteilungsannahmen, wie die Cauchyverteilung, aus theoretischer Sicht sinnvolle Testverfahren. Insbesondere besteht nicht das Problem eines Modellierungsfehlers durch eine unpassende Modellwahl. Eine Diskussion zur numerischen Berechnung der Datentiefe, sowie eine Überlegung zur Konstruktion eines Relevanz-Zweistichproben-Test als Vorzeichentest ergänzen den Abschnitt 2.3. Verteilungsklassen, die in der Simulationsstudie in Kapitel 3 verwendet werden, sowie Kontaminationen von Verteilungen für Robustheitsanalysen, werden in Abschnitt 2.4 dargestellt.

2.1 Allgemeiner Zweistichproben-t-Test

In den Abschnitten 2.1 und 2.2 liegen stets normalverteilte Fehler $E_i \sim \mathcal{N}(0, \sigma^2)$ für $i = 1, ..., M + N$ vor. Aus den Modellgleichungen (1) und (2) ergibt sich, dass

$$X_i \sim \mathcal{N}(\mu_1, \sigma^2) \text{ für } i = 1, ..., M \text{ und}$$
$$Y_j \sim \mathcal{N}(\mu_2, \sigma^2) \text{ für } j = 1, ..., N$$

gelten. Die Normalverteilung wird in vielen statistischen Zusammenhängen zur Modellierung verwendet. Daher ist das Studium von Testverfahren für normalverteilte Stichproben von hoher Interesse.
Zunächst werden schrittweise der Einstichproben- und der Zweistichproben-t-Test jeweils ohne Relevanzparameter δ hergeleitet. Die Rechnungen in den nächsten Abschnitten ergeben sich aus den Standard-t-Tests oder bauen auf diesen auf. Folgende Verteilungsklassen bilden die Grundlage der Theorie (Georgii (2002), S.238f):

Definition 2.1 (χ^2-Verteilung, t-Verteilung).
(a) Eine Zufallsvariable W_N heißt χ^2-verteilt mit N Freiheitsgraden, kurz: $Z \sim \chi^2_N$, falls es unabhängige, identisch verteilte Zufallsvariablen $V_1, ..., V_N \sim \mathcal{N}(0, 1)$ gibt,

$$\textit{sodass } W_N = \sum_{n=1}^{N} V_n^2 \textit{ gilt.}$$

(b) Eine Zufallsvariable T_N heißt t-verteilt mit N Freiheitsgraden, kurz: $T_N \sim t_N$, falls es zwei zueinander stochastisch unabhängige Zufallsvariable $V \sim \mathcal{N}(0, 1)$ und $W_N \sim \chi^2_N$ gibt,

$$\textit{sodass } T_N = \frac{V}{\sqrt{\frac{W_N}{N}}} \textit{ gilt.}$$

Das erste Ziel ist der Vergleich des unbekannten Mittelwerts μ einer Stichprobe X mit einem vorgegeben Mittelwert $\mu_0 \in \mathbb{R}$. Folgendes Hypothesenpaar gibt diese

Fragestellung wieder:

$$H_0 : \mu = \mu_0 \quad \text{gegen} \quad H_1 : \mu \neq \mu_0.$$

Hier sei $\mu = \mu_1$ der unbekannte Lageparameter (in diesem Fall auch der Erwartungswert) der unabhängig, identisch verteilten Zufallsvariablen $X_1, ..., X_M$ zur Stichprobe X. Der unbekannte Lageparameter μ wird mit einem vorgeschlagenen Mittelwert μ_0 verglichen. Ein wichtiger Baustein zur Konstruktion des Einstichproben-t-Tests ist folgender Zusammenhang des Stichprobenmittels und der empirischen Varianz einer Stichprobe unter Normalverteilungsannahme.

Lemma 2.2 (Herleitung der t-Teststatistik).
Seien $V_1, ..., V_N$ unabhängig, identisch verteilte Zufallsvariablen mit $V_n \sim \mathcal{N}(\mu, \sigma^2)$ für $n = 1, ..., N$ und

$$\overline{V.} := \frac{1}{N} \sum_{n=1}^{N} V_n \text{ das Stichprobenmittel über } V = (V_1, ..., V_N)^\top \text{ und}$$

$$\widehat{\sigma}_V^2 := \frac{1}{N-1} \sum_{n=1}^{N} (V_n - \overline{V.})^2 \text{ die empirische Varianz über } V = (V_1, ..., V_N)^\top.$$

Dann gelten folgende Aussagen:

(a) $\frac{N-1}{\sigma^2} \widehat{\sigma}_V^2 \sim \chi^2_{N-1}$,

(b) $\overline{V.}$ *und* $\widehat{\sigma}_V^2$ *sind stochastisch unabhängig,*

(c) $\sqrt{N} \dfrac{\overline{V.} - \mu}{\sqrt{\widehat{\sigma}_V^2}} \sim t_{N-1}$.

Beweis. Die Beweisidee ist an Georgii (2002), S.241f angelehnt. Zunächst werden einige Vorbereitungen getroffen, die dann die Aussagen (a) bis (c) ergeben werden. Man betrachte eine orthogonale Matrix $A \in \mathbb{R}^{N \times N}$ mit folgender letzten Zeile

$$A = \begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \cdots & \frac{1}{\sqrt{N}} \end{pmatrix}.$$

Zum Beispiel erfüllt dies die sogenannte *Helmert-Matrix* mit folgender Gestalt:

$$A := \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{N(N-1)}} & \frac{1}{\sqrt{N(N-1)}} & \frac{1}{\sqrt{N(N-1)}} & \dots & -\frac{N-1}{\sqrt{N(N-1)}} \\ \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \frac{1}{\sqrt{N}} & \dots & \frac{1}{\sqrt{N}} \end{pmatrix}.$$

Die Orthogonalität $AA^\top = I_N$ lässt sich hierbei elementar nachrechnen. Nun wird $V_* := (V_1 - \mu, ..., V_N - \mu)^\top$ als die zentrierte Version von V definiert. Dann hat $V_* \sim \mathcal{N}_N\left(\overrightarrow{0_N}, I_N\sigma^2\right)$ eine N-dimensionale multivariate Normalverteilung. Ferner definiert man die Zufallsvariablen $Z = (Z_1, ..., Z_N)^\top$ durch:

$$Z := A \cdot V_* = A \begin{pmatrix} V_1 - \mu \\ \vdots \\ V_N - \mu \end{pmatrix}.$$

Mit den Rechenregeln für die multivariate Normalverteilungen und der Orthogonalität von A folgt dann:

$$Z \sim \mathcal{N}_N\left(A \cdot \overrightarrow{0_N}, AI_N\sigma^2A^\top\right) = \mathcal{N}_N\left(\overrightarrow{0_N}, AA^\top\sigma^2\right) = \mathcal{N}_N\left(\overrightarrow{0_N}, I_N\sigma^2\right)$$

Aus der Kovarianzmatrix von Z lässt sich ablesen, dass die Zufallsvariablen $Z_1, ..., Z_N$ unkorrelliert sind, was unter Normalverteilung äquivalent zur stochastischen Unabhängigkeit ist (Bauer (2002), S.263). Außerdem fällt auf, dass V_* und Z identisch verteilt sind. Ferner soll $(N-1)\sigma_V^2$ durch Z dargestellt werden. Dabei ist die Anwendung des Verschiebungssatzes nützlich:

$$\begin{aligned} (N-1)\widehat{\sigma}_Y^2 &= \sum_{n=1}^{N}(V_n - \overline{V.})^2 = \sum_{n=1}^{N}(V_n - \mu - (\overline{V.} - \mu))^2 \\ &= \sum_{n=1}^{N}(V_n - \mu)^2 - \left(\sqrt{N}(\overline{V.} - \mu)\right)^2, \qquad (*) \end{aligned}$$

Nun kann man die zwei entstandenen Summanden separat untersuchen. Für den

ersten Summanden ergibt sich die folgende Beziehung:

$$\sum_{n=1}^{N}(V_n-\mu)^2 = V_*^\top V_* = V_*^\top A^\top A V_* = (AV_*)^\top AV_* = Z^\top Z = \sum_{n=1}^{N} Z_n^2 \qquad (**)$$

Hier geht erneut die Orthogonalität von A ein. Der andere Summand kann als Z_N dargestellt werden. Dabei geht die Gestalt der letzten Zeilen der Matrix A ein:

$$\sqrt{N}(\overline{V.}-\mu) = \frac{1}{\sqrt{N}}\sum_{n=1}^{N}(V_n-\mu) = \begin{pmatrix} \frac{1}{\sqrt{N}} & \dots & \frac{1}{\sqrt{N}} \end{pmatrix} \begin{pmatrix} V_1-\mu \\ \vdots \\ V_N-\mu \end{pmatrix} = Z_N \qquad (***)$$

Nun sind alle Bausteine bereit gelegt, um die Aussagen (a) bis (c) zu zeigen:
Für (a) verwendet man die Resultate der Gleichungen $(*)$ bis $(***)$:

$$\begin{aligned} \frac{N-1}{\sigma^2}\widehat{\sigma}_V^2 &\overset{(*)}{=} \frac{1}{\sigma^2}\left(\sum_{n=1}^{N}(V_n-\mu)^2 - \left(\sqrt{N}(\overline{V.}-\mu)\right)^2\right) \\ &\overset{(**)}{=} \frac{1}{\sigma^2}\left(\sum_{n=1}^{N} Z_n^2 - \left(\sqrt{N}(\overline{V.}-\mu)\right)^2\right) \\ &\overset{(***)}{=} \frac{1}{\sigma^2}\left(\sum_{n=1}^{N} Z_n^2 - Z_N^2\right) \\ &= \sum_{n=1}^{N-1}\left(\frac{Z_n}{\sigma}\right)^2 \sim \chi^2_{N-1} \end{aligned}$$

In der letzten Zeile werden die Unabhängigkeit der Zufallsvariablen $Z_1, ..., Z_{N-1}$ und $Z_n \sim \mathcal{N}(0,\sigma^2)$ für $n = 1, ..., N-1$ ausgenutzt. Somit liegt nach Definition 2.1(a) eine χ^2-Verteilung mit $N-1$ Freiheitsgraden vor, womit (a) gezeigt ist.
Die Aussage (b) zeigt man, indem man $\widehat{\sigma}_V^2$ und $\overline{V.}$ umschreibt. Mit der Rechnung in (a) ergibt sich für $\widehat{\sigma}_V^2$ die Gestalt:

$$\widehat{\sigma}_V^2 = \frac{\sigma^2}{N-1}\sum_{n=1}^{N-1} Z_n^2$$

und $(***)$ liefert entsprechend:

$$\overline{V.} = \frac{1}{\sqrt{N}} Z_N + \mu.$$

Mit dieser Darstellung folgt die Unabhängigkeit von $\widehat{\sigma}_V^2$ und $\overline{V.}$ dann direkt aus der Unabhängigkeit der Zufallsvariablen $Z_1, ..., Z_{N-1}$ und Z_N, was (b) zeigt.
Zu (c) erweitert man zunächst die zu untersuchende Zufallsvariable mit $\frac{1}{\sigma}$:

$$\sqrt{N}\,\frac{\overline{V.} - \mu}{\sqrt{\widehat{\sigma}_V^2}} = \frac{\frac{\sqrt{N}}{\sigma}(\overline{V.} - \mu)}{\frac{1}{\sigma}\sqrt{\widehat{\sigma}_V^2}}$$

und betrachtet Zähler und Nenner der rechten Seite getrennt. Der Zähler hat eine $\mathcal{N}(0,1)$-Verteilung, da $\overline{V.} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$ gilt. Den Nenner schreibt man um:

$$\frac{1}{\sigma}\sqrt{\widehat{\sigma}_V^2} = \sqrt{\frac{\frac{(N-1)\widehat{\sigma}_V^2}{\sigma^2}}{N-1}}.$$

Nach der Rechnung aus (a) ist bekannt, dass $\frac{(N-1)\widehat{\sigma}_V^2}{\sigma^2} \sim \chi^2_{N-1}$ gilt. Außerdem folgt aus (b) die Unabhängigkeit von Zähler und Nenner. Die zu untersuchende Zufallsvariable hat also nach Definition 2.1(b) die Form einer t-verteilten Zufallsvariable mit $N-1$ Freiheitsgraden:

$$\sqrt{N}\,\frac{\overline{V.} - \mu}{\sqrt{\widehat{\sigma}_V^2}} \sim t_{N-1}.$$

Damit ist auch die letzte Aussage (c) gezeigt. □

Mithilfe des Lemmas liegen alle Hilfsmittel bereit den Einstichproben-t-Test zu formulieren und zu zeigen, dass er ein vorgegebenes Signifikanzniveau unter der Nullhypothese einhält (siehe auch Sachs und Hedderich (2015), S.478f).

Satz 2.3 (Einstichproben-t-Test).
Seien $V_1, ..., V_N$ unabhängig, identisch verteilte Zufallsvariablen mit $V_n \sim \mathcal{N}(\mu, \sigma^2)$

mit unbekanntem Parameter $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ *und gegebenem* $\mu_0 \in \mathbb{R}$. *Ferner sei*

$$T(V) = T(V_1, ..., V_N) = \sqrt{N}\frac{\overline{V.} - \mu_0}{\sqrt{\widehat{\sigma}_V^2}}$$

die Teststatistik. Dann liefern folgende Entscheidungsregeln φ *zu den gegebenen Hypothesenpaaren jeweils Tests zum Niveau* α*:*

(a) $\varphi(v_1, ..., v_N) = \mathbb{1}_{\{|T(v)| > t_{N-1,1-\frac{\alpha}{2}}\}}$ *mit* $H_0 : \mu = \mu_0$ *gegen* $H_1 : \mu \neq \mu_0$,

(b) $\varphi(v_1, ..., v_N) = \mathbb{1}_{\{T(v) > t_{N-1,1-\alpha}\}}$ *mit* $H_0 : \mu \leq \mu_0$ *gegen* $H_1 : \mu > \mu_0$,

(c) $\varphi(v_1, ..., v_N) = \mathbb{1}_{\{T(v) < t_{N-1,\alpha}\}}$ *mit* $H_0 : \mu \geq \mu_0$ *gegen* $H_1 : \mu < \mu_0$,

wobei $t_{N,\alpha}$ *das* α*-Quantil der* t*-Verteilung mit* N *Freiheitsgraden sei.*

Beweis. Beim Testverfahren (a) gilt unter der Nullhypothese $\mu = \mu_0$. Daher gilt $V_n \sim \mathcal{N}(\mu_0, \sigma^2)$ für alle $n = 1, ..., N$. Lemma 2.2 impliziert außerdem, dass die Teststatistik unter der Nullhypothese t-verteilt mit $N - 1$ Freiheitsgraden ist:

$$T(V) = \sqrt{N}\frac{\overline{V.} - \mu_0}{\sqrt{\widehat{\sigma}(V)^2}} \sim t_{N-1}.$$

Nun wird bewiesen, dass unter der Nullhypothese die Wahrscheinlichkeit diese abzulehnen höchstens α ist.

$$\begin{aligned} P_{\mu_0}(\varphi(V) = 1) &= P_{\mu_0}(|T(V)| > t_{N-1,1-\frac{\alpha}{2}}) \\ &= P_{\mu_0}\big(\{T(V) < -t_{N-1,1-\frac{\alpha}{2}}\} \cup \{T(V) > t_{N-1,1-\frac{\alpha}{2}}\}\big) \\ &= P_{\mu_0}(T(V) < t_{N-1,\frac{\alpha}{2}}) + P_{\mu_0}(T(V) > t_{N-1,1-\frac{\alpha}{2}}) \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha. \end{aligned}$$

Man beachte, dass die Ereignisse in der zweiten Zeile disjunkt sind. Somit verhält sich die Wahrscheinlichkeit ihrer Vereinigung additiv.
Nun erfolgt der Beweis von (b). Sei μ aus dem Annahmebereich mit der Eigenschaft

$\mu \leq \mu_0$. Insbesondere gilt dann $\mu_0 - \mu \geq 0$. Man berechne nun den Fehler erster Art:

$$\begin{aligned} P_\mu(\varphi(V) = 1) &= P_\mu\left(T(V) > t_{N-1,1-\alpha}\right) \\ &= P_\mu\left(\sqrt{N}\frac{\overline{V.} - \mu_0}{\sqrt{\widehat{\sigma}_V^2}} > t_{N-1,1-\alpha}\right) \\ &= P_\mu\left(\sqrt{N}\frac{\overline{V.} - \mu}{\sqrt{\widehat{\sigma}_V^2}} > t_{N-1,1-\alpha} + \underbrace{\frac{\mu_0 - \mu}{\sqrt{\widehat{\sigma}_V^2}}}_{\geq 0}\right) \end{aligned}$$

An dieser Stelle soll ausgenutzt werden, dass $\mu_0 - \mu \geq 0$ gilt, indem man in der letzten Zeile den nicht-negativen Bruch weglässt. Dadurch wird die untere Schranke der Ungleichung kleiner, wodurch die betrachtete Ereignismenge und damit auch die Wahrscheinlichkeit vergrößert wird. Insgesamt folgt also

$$P_\mu(\varphi(V) = 1) \leq P_\mu\left(\sqrt{N}\frac{\overline{V.} - \mu}{\sqrt{\widehat{\sigma}_V^2}} > t_{N-1,1-\alpha}\right) = \alpha.$$

Die letzte Gleichheit folgt aus Lemma 2.2. Damit hält auch das zweite Testverfahren das Signifikanzniveau ein. Der Beweis für Test (c) geht analog. □

Nun folgt die Konstruktion des Zweistichproben-t-Tests aus den bisherigen Ergebnissen. Die Grundidee dieses Testverfahrens ist der Vergleich von den unbekannten Mittelwerten zweier Stichproben, indem ihre Abweichungen durch die beobachten Stichprobenmittel bestimmt und passend standardisiert werden. Für ein festes, bekanntes $\Delta \in \mathbb{R}$, das der vorgeschlagenen Differenz der Mittelwerte entspricht, wird folgendes Hypothesenpaar formuliert:

$$H_0 : \mu_1 - \mu_2 = \Delta \quad \text{gegen} \quad H_1 : \mu_1 - \mu_2 \neq \Delta. \tag{4}$$

In den meisten Fällen ist man an einem generellen Unterschied der Mittelwerte interessiert und setzt dann $\Delta = 0$. Dann lässt sich das Hypothesenpaar in folgender anschaulichen Weise umformen:

$$H_0 : \mu_1 = \mu_2 \quad \text{gegen} \quad H_1 : \mu_1 \neq \mu_2. \tag{5}$$

Erneut wird Lemma 2.2 benötigt, um für einen Zweistichproben-t-Test zu zeigen, dass er ein vorgegebenes Signifikanzniveau einhält. Außerdem müssen die geschätzten Varianzen beider Stichproben geeignet kombiniert werden. Zu den Zufallsvariablen $X_1, ..., X_M$ bzw. $Y_1, ..., Y_N$ wird jeweils eine Realisation für die Stichprobe betrachtet. Die empirischen Varianzen beider Stichproben $\widehat{\sigma}_X^2$ und $\widehat{\sigma}_Y^2$ aus dem Lemma 2.2 werden verwendet, um die gepaarte empirische Varianz $\widehat{\sigma}_{X,Y}^2$ zu erhalten:

$$\widehat{\sigma}_{X,Y}^2 := \frac{1}{M+N-2}\left((M-1)\widehat{\sigma}_X^2 + (N-1)\widehat{\sigma}_Y^2\right)$$

Die kombinierte empirische Varianz lässt sich im Wesentlichen durch die Unabhängigkeit der beiden Stichproben motivieren. Zunächst müssen die Vorfaktoren $\frac{1}{M-1}$ bzw. $\frac{1}{N-1}$ eliminiert werden damit die gesamte empirische Varianz beider Stichproben sinnvoll zusammengefasst werden kann. Die Varianz der Summe von Zufallsvariablen entspricht der Summe der Varianzen der einzelnen Zufallsvariablen unter Unabhängigkeit. Nach diesem Konzept lässt sich die Berechnung der gepaarten empirischen Varianz veranschaulichen. Die Zahl der Freiheitsgrade beruht darauf, dass in den Stichproben bereits $(M-1)$ und $(N-1)$ Freiheitsgrade vorliegen. Dies ist zu beachten damit die Teststatistik unter der Nullhypothese t-verteilt ist. Der Faktor $K := \sqrt{\frac{MN}{M+N}}$ wird im Folgenden bei allen Zweistichproben-t-Tests und den Relevanz-Tests als Abkürzung verwendet (Sachs und Hedderich (2015), S.508f).

Satz 2.4 (Zweistichproben-t-Test).
Seien $V = (V_1, ..., V_M), W = (W_1, ..., W_N)$ unabhängige Zufallsvariablen mit $V_m \sim \mathcal{N}(\mu_1, \sigma^2)$ für $m = 1, ..., M$ und $W_n \sim \mathcal{N}(\mu_2, \sigma^2)$ für $n = 1, ..., N$ mit unbekannten Parametern $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$ und gegebenem $\Delta \in \mathbb{R}$. Ferner sei

$$T(V, W) = K\,\frac{\overline{V.} - \overline{W.} - \Delta}{\sqrt{\widehat{\sigma}_{V,W}^2}}$$

die Teststatistik. Folgende Entscheidungsregeln $\varphi(v, w) = \varphi(v_1, ..., v_M, w_1, ..., w_N)$ liefern zu den gegebenen Hypothesenpaaren dann jeweils Tests zum Niveau α:

(a) $\varphi(v, w) = \mathbb{1}_{\{|T(v,w)| > t_{M+N-2,1-\frac{\alpha}{2}}\}}$ mit $H_0 : \mu_1 - \mu_2 = \Delta$ gegen $H_1 : \mu_1 - \mu_2 \neq \Delta$,

(b) $\varphi(v,w) = \mathbb{1}_{\{T(v,w) > t_{M+N-2,1-\alpha}\}}$ *mit* $H_0 : \mu_1 - \mu_2 \leq \Delta$ *gegen* $H_1 : \mu_1 - \mu_2 > \Delta$,

(c) $\varphi(v,w) = \mathbb{1}_{\{T(v,w) < t_{M+N-2,\alpha}\}}$ *mit* $H_0 : \mu_1 - \mu_2 \geq \Delta$ *gegen* $H_1 : \mu_1 - \mu_2 < \Delta$,

wobei $t_{N,\alpha}$ *das* α*-Quantil der* t*-Verteilung mit* N *Freiheitsgraden sei.*

Beweis. Zunächst wird die Aussage (a) bewiesen. Die Verteilung der dort angegebenen Zufallsvariable wird schrittweise bestimmt. Folgende Verteilungseigenschaft ergibt sich aus der Unabhängigkeit der Zufallsvariablen

$$\overline{V.} - \overline{W.} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{M} + \frac{\sigma^2}{N}\right) = \mathcal{N}\left(\mu_1 - \mu_2, \sigma^2 \frac{M+N}{MN}\right).$$

Es sei an die Abkürzung $K := \sqrt{\dfrac{MN}{M+N}}$ zur Normierung der Varianz erinnert. Erweitert man die Teststatistik mit $\frac{1}{\sigma}$ ergibt sich für ihren Zähler

$$\frac{K}{\sigma}(\overline{V.} - \overline{W.} - (\mu_1 - \mu_2)) \sim \mathcal{N}(0,1).$$

Andererseits folgt für den Nenner der Teststatistik $T(V,W)$ nach Erweitern:

$$\frac{\widehat{\sigma}^2_{V,W}}{\sigma} = \frac{1}{M+N-2} \underbrace{\left(\sum_{m=1}^{M}\left(\frac{V_m - \overline{V.}}{\sigma}\right)^2 + \sum_{n=1}^{N}\left(\frac{W_m - \overline{W.}}{\sigma}\right)^2\right)}_{\sim \chi^2_{M+N-2}},$$

wegen Lemma 2.2(a) durch die Unabhängigkeit von V und W. Schließlich liefert Lemma 2.2(c) die t-Verteilung der Teststatistik $T(V,W)$ unter der Nullhypothese. Zu (b): Man betrachte dazu Parameter $\mu = (\mu_1, \mu_2)$ unter der Nullhypothese mit $\mu_1 - \mu_2 \leq \Delta$ bzw. $\Delta - (\mu_1 - \mu_2) \geq 0$. Für den Fehler erster Art gilt:

$$\begin{aligned}
P_\mu(T(V,W) > t_{M+N-2;1-\alpha}) &= P_\mu\left(K\,\frac{\overline{V.} - \overline{W.} - \Delta}{\sqrt{\widehat{\sigma}^2_{V,W}}} > t_{M+N-2;1-\alpha}\right) \\
&= P_\mu\left(K\,\frac{\overline{V.} - \overline{W.} - (\mu_1 - \mu_2)}{\sqrt{\widehat{\sigma}^2_{V,W}}} > t_{M+N-2;1-\alpha} + \underbrace{K\,\frac{\Delta - (\mu_1 - \mu_2)}{\sqrt{\widehat{\sigma}^2_{V,W}}}}_{\geq 0}\right) \\
&\leq P_\mu\left(K\,\frac{\overline{V.} - \overline{W.} - (\mu_1 - \mu_2)}{\sqrt{\widehat{\sigma}^2_{V,W}}} > t_{M+N-2;1-\alpha}\right) = \alpha.
\end{aligned}$$

Der Beweis von (c) geht analog. □

2.2 Zweistichproben-Relevanz-t-Tests

Nun werden drei Testverfahren vorgestellt, die als Relevanz-t-Tests bezeichnet werden, da sie aus den t-Tests in Abschnitt 2.1 konstruiert werden. Zur Erinnerung wird das Hypothesenpaar des Relevanz-Tests mit Relevanzparameter $\delta \leq 0$ angegeben:

$$H_0 : |\mu_1 - \mu_2| \leq \delta \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > \delta \tag{H}$$

2.2.1 Konfidenzintervall-Exlusion

Die erste Methode zur Konstruktion eines Relevanz-t-Tests beruht auf die Verwendung eines Konfidenzintervalls für $\Delta = \mu_1 - \mu_2$ zum Konfidenzniveau $1 - \alpha$. Überdeckt die Bereichsschätzung nicht den Nicht-Relevanzbereich $[-\delta, \delta]$, so kann die Nullhypothese (H) zum Niveau α abgelehnt werden. Letztere Aussage soll nun im Folgenden nachgewiesen werden. Das verwendete Konfidenzintervall für $\Delta = \mu_1 - \mu_2$ wird in Satz 2.5 hergeleitet.
Der hier gewählte Ansatz entspricht einer analogen Überlegung, die zur Konstruktion von Äquivalenztests mittels Konfidenzintervall-Inklusion verwendet wird (Schuhmacher und Schulgen (2002), S.109f). Bei diesem Ansatz muss $[-\delta, \delta]$ vom Konfidenzintervall (umgekehrt zu Relevanz-Tests) eingeschlossen werden, um die Nullhypothese abzulehnen. Analog wird folgender Ansatz zur Konstruktion des Relevanz-Tests daher Relevanz-Tests mittels *Konfidenzintervall-Exklusion* bezeichnet.

Satz 2.5 (Konfidenzintervall des Zweistichproben-t-Test für Δ).
Seien $V_1, ..., V_M, W_1, ..., W_N$ unabhängige Zufallsvariablen mit $V_m \sim \mathcal{N}(\mu_1, \sigma^2)$ für $m = 1, ..., M$ und $W_n \sim \mathcal{N}(\mu_2, \sigma^2)$ für $n = 1, ..., N$ mit unbekannten Parametern $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$. Dann erhält man für $\Delta = \mu_1 - \mu_2$ ein $(1 - \alpha)$-Konfidenzintervall durch

$$\left[\overline{V.} - \overline{W.} - t_{M+N-2,1-\frac{\alpha}{2}} \frac{\widehat{\sigma}_{V,W}}{K}, \overline{V.} - \overline{W.} + t_{M+N-2,1-\frac{\alpha}{2}} \frac{\widehat{\sigma}_{V,W}}{K} \right],$$

wobei $\widehat{\sigma}_{V,W} = \sqrt{\widehat{\sigma}^2_{V,W}}$ *ist.*

Beweis. Sei $\mu = (\mu_1, \mu_2)$ der wahre Parameter mit $\mu_1 - \mu_2 = \Delta$. Zu zeigen ist, dass obiges Konfidenzintervall das Konfidenzniveau $1-\alpha$ für Δ einhält. Dazu werden die Ungleichung soweit umgeformt, sodass die Teststatistik aus Satz 2.4 vorliegt.

$$\begin{aligned}
&P_\mu\left(\overline{V.} - \overline{W.} - t_{M+N-2,1-\frac{\alpha}{2}}\frac{\widehat{\sigma}_{V,W}}{K} \leq \Delta \leq \overline{V.} - \overline{W.} + t_{M+N-2,1-\frac{\alpha}{2}}\frac{\widehat{\sigma}_{V,W}}{K}\right)\\
&= P_\mu\left(-t_{M+N-2,1-\frac{\alpha}{2}} \leq -K\,\frac{\overline{V.} - \overline{W.} - \Delta}{\widehat{\sigma}_{V,W}} \leq t_{M+N-2,1-\frac{\alpha}{2}}\right)\\
&= P_\mu\left(t_{M+N-2,1-\frac{\alpha}{2}} \geq -K\,\frac{\overline{V.} - \overline{W.} - \Delta}{\widehat{\sigma}_{V,W}} \geq t_{M+N-2,\frac{\alpha}{2}}\right)\\
&= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.
\end{aligned}$$

Man beachte $-t_{M+N-2,1-\frac{\alpha}{2}} = t_{M+N-2,\frac{\alpha}{2}}$ und dass in der letzten Zeile $T(X,Y) \sim t_{M+N-2}$ ausgenutzt wird. □

Das hergeleitete Konfidenzintervall erlaubt nun die Konstruktion eines Relevanz-Tests zum Niveau α.

Satz 2.6 (Zweistichproben-Relevanz-t-Test durch Konfidenzintervalle). *Seien* $V_1, ..., V_M, W_1, ..., W_N$ *unabhängige Zufallsvariablen mit* $V_m \sim \mathcal{N}(\mu_1, \sigma^2)$ *für* $m = 1, ..., M$ *und* $W_n \sim \mathcal{N}(\mu_2, \sigma^2)$ *für* $n = 1, ..., N$ *mit unbekannten Parametern* $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$. *Ferner seien* $U_{V,W}$ *bzw.* $O_{V,W}$ *die untere bzw. obere Grenze des Konfidenzintervalls aus Satz 2.5*

$$\begin{aligned}
U_{V,W} &:= \overline{V.} - \overline{W.} - t_{M+N-2,1-\alpha/2}\frac{\widehat{\sigma}_{V,W}}{K},\\
O_{V,W} &:= \overline{V.} - \overline{W.} + t_{M+N-2,1-\alpha/2}\frac{\widehat{\sigma}_{V,W}}{K}.
\end{aligned}$$

zum Konfidenzniveau $1-\alpha$. *Dann liefert die Entscheidungsregel*

$$\varphi(v,w) = \varphi(v_1, ..., v_M; w_1, ..., w_N) = \mathbb{1}_{\left\{\{U_{v,w} > \delta\} \cup \{O_{v,w} < -\delta\}\right\}}$$

für das Hypothesenpaar $H_0 : |\mu_1 - \mu_2| \leq \delta$ *gegen* $H_1 : |\mu_1 - \mu_2| > \delta$ *einen Test zum Niveau* α.

Beweis. Sei $\mu = (\mu_1, \mu_2)$ aus dem Annahmebereich, d.h. $|\mu_1 - \mu_2| \leq \delta$ gilt. Es muss gezeigt werden, dass der Fehler erster Art für alle μ durch α beschränkt ist:

$$\begin{aligned} P_\mu(\varphi(V,W) = 1) &= P_\mu(\{U_{V,W} > \delta\} \cup \{O_{V,W} < -\delta\}) \\ &= P_\mu(U_{V,W} > \delta) + P_\mu(O_{V,W} < -\delta) \end{aligned}$$

Da der Nicht-Relevanzbereich $[-\delta, \delta]$ entweder durch die untere Schranke der Bereichsschätzung überschritten oder durch ihre obere Schranke unterschritten wird, kann im letzten Schritt die Disjunktheit der Ereignisse verwendet werden. Man zeige nun für beide Wahrscheinlichkeiten, dass sie durch $\frac{\alpha}{2}$ beschränkt sind.

$$\begin{aligned} &P_\mu(U_{V,W} > \delta) = P_\mu\left(\overline{V.} - \overline{W.} - t_{M+N-2,1-\alpha/2}\frac{\widehat{\sigma}_{V,W}}{K} > \delta\right) \\ &= P_\mu\left(K\frac{\overline{V.} - \overline{W.} - \delta}{\widehat{\sigma}_{V,W}} > t_{M+N-2,1-\alpha/2}\right) \\ &= P_\mu\Bigg(K\Bigg(\frac{\overline{V.} - \overline{W.} - (\mu_1 - \mu_2)}{\widehat{\sigma}_{V,W}} + \underbrace{\frac{(\mu_1 - \mu_2) - \delta}{\widehat{\sigma}_{V,W}}}_{\leq 0}\Bigg) > t_{M+N-2,1-\alpha/2}\Bigg). \end{aligned}$$

Aus der Annahme $\mu_1 - \mu_2 \leq \delta$ folgt $\mu_1 - \mu_2 - \delta \leq 0$. Entfernt man den Term $\mu_1 - \mu_2 - \delta$ aus dem Zähler in der letzten Wahrscheinlichkeit, so wird die Ereignismenge und damit die Wahrscheinlichkeit vergrößert. Ferner besitzt der restliche Ausdruck nach Satz 2.4 folgende Verteilung unter der Nullhypothese:

$$K\frac{\overline{V.} - \overline{W.} - (\mu_1 - \mu_2)}{\widehat{\sigma}_{V,W}} \sim t_{M+N-2}.$$

Daraus ergibt sich das gewünschte Teilresultat:

$$P_\mu(U_{V,W}) \leq P_\mu\left(K\frac{\overline{V.} - \overline{W.} - (\mu_1 - \mu_2)}{\widehat{\sigma}_{V,W}} > t_{M+N-2,1-\alpha/2}\right) = \frac{\alpha}{2}.$$

Die Argumentation für die andere Wahrscheinlichkeit verläuft analog:

$$P_\mu(O_{V,W} < -\delta) = P_\mu\left(K\left(\frac{\overline{V.} - \overline{W.} - (\mu_2 - \mu_1)}{\widehat{\sigma}_{V,W}} + \frac{(\mu_2 - \mu_1) + \delta}{\widehat{\sigma}_{V,W}}\right) < t_{M+N-2,\alpha/2}\right)$$

Nach Annahme gilt $\mu_2 - \mu_1 \geq -\delta$, was $\mu_2 - \mu_1 + \delta \geq 0$ impliziert. Entfernt man

diesen Term im Zähler, so wird analog wie oben die Wahrscheinlichkeit vergrößert:

$$P_\mu\left(O_{V,W} < -\delta\right) \leq P_\mu\left(K\frac{\overline{V.} - \overline{W.} - (\mu_2 - \mu_1)}{\widehat{\sigma}_{V,W}} < t_{M+N-2,\alpha/2}\right) = \frac{\alpha}{2}.$$

Dabei geht hier erneut Satz 2.4 ein. Die Zusammensetzung beider Teilresultate liefert, dass der Fehler erster Art durch α beschränkt ist. □

Die Grenzen des Konfidenzintervalls aus Satz 2.5 können für die Konstruktion eines Relevanz-Tests mit Konfidenzintervall-Exklusion verallgemeinert werden.
Für $\alpha_1, \alpha_2 \in (0, \alpha)$ seien

$$U_{V,W}(\alpha_1) := \overline{V.} - \overline{W.} - t_{M+N-2,1-\alpha_1}\frac{\widehat{\sigma}_{V,W}}{K} \text{ und}$$
$$O_{V,W}(\alpha_2) := \overline{V.} - \overline{W.} + t_{M+N-2,1-\alpha_2}\frac{\widehat{\sigma}_{V,W}}{K},$$

wobei $\alpha_1 + \alpha_2 = \alpha$ gelte. Ersetzt man in Satz 2.6 die Grenzen $U_{V,W}$ und $O_{V,W}$ durch $U_{V,W}(\alpha_1)$ und $O_{V,W}(\alpha_2)$, so erhält man ebenfalls einen Relevanz-Test zum Niveau α. Der Beweis dazu geht vollständig analog. Im Fall von Satz 2.6 sind $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$. Dadurch lassen sich die beiden möglich vorkommenden Lageunterschiede unterschiedlich stark gewichten. Für $\alpha_1 \to 0$ bzw. $\alpha_2 \to 0$ ergibt sich ein zu einem der einseitigen t-Tests aus Satz 2.3 äquivalenter Test.

Zum Abschluss zur ersten Variante des Relevanz-t-Tests soll der Fall für $\delta = 0$ betrachtet werden. Die Entscheidungsregel des Tests 2.6 besitzt dann folgende Gestalt

$$\varphi(v, w) = \mathbb{1}_{\left\{\{U_{v,w}>0\}\cup\{O_{v,w}<0\}\right\}} = \mathbb{1}_{\left\{0\notin[U_{v,w},O_{v,w}]\right\}}.$$

Diese Entscheidungsregel untersucht die Fragestellung, ob der Parameter $\Delta = 0$ im realisierten Konfidenzintervall aus Satz 2.5 enthalten ist. Das wiederum entspricht dem Testproblem aus Satz 2.4 geschrieben als *duales Testproblem* mittels Konfidenzintervall (Czado und Schmidt (2011), S.157f). Der Relevanz-Test ist für den Fall $\delta = 0$ also entartet und entspricht einem üblichen Zweistichproben-t-Test ohne Relevanzbereiche.

2.2.2 Multiple Hypothesen

Ein weiterer elementarer Ansatz ist die Auflösung der Hypothesen in zwei Teilhypothesen, die man dann jeweils separat untersucht:

$$\mathrm{H}_0 : |\mu_1 - \mu_2| \leq \delta \quad \text{gegen} \quad \mathrm{H}_1 : |\mu_1 - \mu_2| > \delta \tag{H}$$

$$\mathrm{H}_0^{(1)} : \mu_1 - \mu_2 \leq \delta \quad \text{gegen} \quad \mathrm{H}_1^{(1)} : \mu_1 - \mu_2 > \delta \tag{H$^{(1)}$}$$

$$\mathrm{H}_0^{(2)} : \mu_2 - \mu_1 \leq \delta \quad \text{gegen} \quad \mathrm{H}_1^{(2)} : \mu_2 - \mu_1 > \delta \tag{H$^{(2)}$}$$

Die beiden Teilhypothesenpaare $\mathrm{H}^{(i)}$ für $i = 1, 2$ können jeweils mit einem einseitigen Zweistichproben-t-Test mit $\Delta = \delta$ getestet werden. Dadurch wird aus einem zweiseitigen Hypothesenpaar ein multiples Testproblem mit jeweils zwei einseitigen Hypothesen generiert.
Man lehnt die *globale Nullhypothese* H_0 ab, falls $\mathrm{H}_0^{(1)}$ oder $\mathrm{H}_0^{(2)}$ abgelehnt werden. Das Signifikanzniveau wird im multiplen Testverfahren jeweils adjustiert, um einen inflationären Effekt auf den Fehler erster Art zu verhindern. Das *globale Signifikanzniveau* α, das für das gesamte Testverfahren (H) eingehalten soll, muss in den Teilhypothesen ($\mathrm{H}^{(i)}$) für $i = 1, 2$ verkleinert werden. Das jeweilige adjustierte Signifikanzniveau $\alpha^{(i)}$ zur Teilhypothese ($\mathrm{H}^{(i)}$) heißt lokales Signifikanzniveau.
Ein Weg zur Adjustierung ermöglicht die Bonferroni-Methode. Dabei werden das lokale Signifikanzniveau durch die Anzahl aller durchgeführten Tests dividiert; hier also durch $\alpha^{(i)} = \frac{\alpha}{2}$ für $i = 1, 2$ (siehe auch Sachs und Hedderich (2015), S.568ff).

Satz 2.7 (Multipler Zweistichproben-Relevanz-t-Test).
Seien $V_1, ..., V_M, W_1, ..., W_N$ unabhängige Zufallsvariablen mit $V_m \sim \mathcal{N}(\mu_1, \sigma^2)$ für $m = 1, ..., M$ und $W_n \sim \mathcal{N}(\mu_2, \sigma^2)$ für $n = 1, ..., N$ mit unbekannten Parametern $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$. Ferner seien

$$T^{(1)}(V, W) = K \, \frac{\overline{V.} - \overline{W.} - \delta}{\widehat{\sigma}_{V,W}}$$

$$T^{(2)}(V, W) = K \, \frac{\overline{W.} - \overline{V.} - \delta}{\widehat{\sigma}_{V,W}}$$

die Teststatistiken für das Hypothesenpaar ($H^{(1)}$) bzw. ($H^{(2)}$). Dann liefert die Ent-

scheidungsregel

$$\varphi(v,w) = \varphi(v_1,...,v_M,w_1,...,w_N) = \mathbb{1}_{\left\{\{T^{(1)}(v,w) > t_{M+N-2,1-\frac{\alpha}{2}}\} \cup \{T^{(2)}(v,w) > t_{M+N-2,1-\frac{\alpha}{2}}\}\right\}}$$

für das Hypothesenpaar $H_0 : |\mu_1 - \mu_2| \leq \delta$ *gegen* $H_1 : |\mu_1 - \mu_2| > \delta$ *einen Test zum Niveau* α*.*

Beweis. Sei $\mu = (\mu_1, \mu_2)$ ein Parameter mit der Eigenschaft $|\mu_1 - \mu_2| \leq \delta$. Man schätze nun den Fehler erster Art gegen α ab:

$$\begin{aligned} P_\mu(\varphi(V,W) = 1) &= P_\mu(\{T^{(1)}(V,W) > t_{M+N-2,1-\frac{\alpha}{2}}\} \cup \{T^{(2)}(V,W) > t_{M+N-2,1-\frac{\alpha}{2}}\}) \\ &\leq \underbrace{P_\mu(T^{(1)}(V;W) > t_{M+N-2,1-\frac{\alpha}{2}})}_{\leq \frac{\alpha}{2}} + \underbrace{P_\mu(T^{(2)}(V,W) > t_{M+N-2,1-\frac{\alpha}{2}})}_{\leq \frac{\alpha}{2}}. \end{aligned}$$

Die beiden Summanden entsprechen exakt den Summanden aus dem Beweis von Satz 2.4(b), da die Untersuchung der Teilhypothesen genau diesem Test entsprechen. Daraus ergibt sich die Beschränktheit durch α. □

Die Bonferroni-Methode ist nicht die einzige Möglichkeit zur Adjustierung des Signifikanzniveaus beim multiplen Testen. Man kann beide Teilhypothese auch durch unterschiedliche lokale Signifikanzniveaus $\alpha_1, \alpha_2 \in (0, \alpha)$ mit $\alpha_1 + \alpha_2 = \alpha$ ersetzen. Vergleicht man in Satz 2.7 die Teststatistiken $T^{(i)}$ statt mit dem $1 - \frac{\alpha}{2}$-Quantil mit dem $1 - \alpha_i$-Quantil für $i = 1, 2$, so ergibt sich auch ein Relevanz-Test zum Niveau α. Dies kann man analog wie in Satz 2.7 beweisen.
Der Relevanz-t-Tests aus Abschnitt 2.2.1 mit Konfidenzintervall-Exklusion und 2.2.2 durch multiple Hypothesen weisen Ähnlichkeiten auf; beispielsweise besitzen beide Entscheidungsregeln eine disjunktive Form (Oder-Verknüpfung). In der Tat kann man leicht die Äquivalenz der Testverfahren nachrechnen. Die Äquivalenz von zwei Tests bedeutet, dass die Testverfahren immer zu gleichen Entscheidungen führen. Um das einzusehen, formt man die Entscheidungsregel des Testverfahrens aus Satz 2.6 in die Entscheidungsregel in Satz 2.7 um. Exemplarisch wird dies für den ersten

Fall gezeigt.

$$\begin{aligned} & U_{V,W} > \delta \\ \Leftrightarrow\ & \overline{V.} - \overline{W.} - t_{M+N-2,1-\frac{\alpha}{2}} \frac{\widehat{\sigma}_{V,W}}{K} > \delta \\ \Leftrightarrow\ & K \frac{\overline{V.} - \overline{W.} - \delta}{\widehat{\sigma}_{V,W}} > t_{M+N-2,1-\frac{\alpha}{2}} \\ \Leftrightarrow\ & T^{(1)}(V,W) > t_{M+N-2,1-\frac{\alpha}{2}} \end{aligned}$$

Die Äquivalenz: $O_{V,W} < -\delta \Leftrightarrow T^{(2)}(V,W) > t_{M+N-2,1-\frac{\alpha}{2}}$ zeigt man analog. Daraus folgt die Äquivalenz der Testverfahren in 2.2.1 und 2.2.2.

Abschließend soll der entartete Fall für $\delta = 0$ untersucht werden. Dieser ergibt dann ebenso einen Zweistichproben-t-Test ohne Relevanzbereich mit Differenzparameter $\Delta = 0$. Das sieht man ein, indem man in die Teststatistiken $T^{(1)}(V,W)$ und $T^{(2)}(V,W)$ den Relevanzparameter $\delta = 0$ setzt. Die Entscheidungsregel aus Satz 2.7

$$\varphi(v,w) = \mathbb{1}_{\left\{\{T^{(1)}(v,w) > t_{M+N-2,1-\frac{\alpha}{2}}\} \cup \{T^{(2)}(v,w) > t_{M+N-2,1-\frac{\alpha}{2}}\}\right\}}$$

ist genau dann 1, falls

$$T^{(1)}(v,w) > t_{M+N-2,1-\frac{\alpha}{2}} \text{ oder } T^{(2)}(v;w) > t_{M+N-2,1-\frac{\alpha}{2}} \text{ gilt.}$$

Es gilt $-T^{(1)}(v,w) = T^{(2)}(v,w)$ wegen $\delta = 0$. Die Aussage ist dann äquivalent zu

$$\begin{aligned} & T^{(1)}(v,w) > t_{M+N-2,1-\frac{\alpha}{2}} \text{ oder } -T^{(1)}(v;w) > t_{M+N-2,\frac{\alpha}{2}} \text{ bzw.} \\ & |T^{(1)}(v,w)| > t_{M+N-2,1-\frac{\alpha}{2}} \end{aligned}$$

Das entspricht genau der Entscheidungsregel aus Satz 2.4(a). Alternativ hätte man dies mit der Äquivalenz zum Test aus Satz 2.2.1 begründen können.

2.2.3 Verwendung der dezentralen t-Verteilung

Für den letzten Ansatz zur Gewinnung eines Zweistichproben-Relevanz-t-Tests wird die *dezentrale t-Verteilung* eingeführt (Sachs und Hedderich (2015), S.289f).

Definition 2.8 (Dezentrale t-Verteilung).
Eine Zufallsvariable Z_N heißt dezentral-t-verteilt mit N Freiheitsgraden und Nichtzentralitätsparameter δ, kurz: $Z_N \sim t_N(\delta)$, falls es zwei zueinander stochastisch unabhängige Zufallsvariable $V \sim \mathcal{N}(\delta, 1)$ und $W_N \sim \chi^2_N$ gibt, sodass

$$T_N = \frac{V}{\sqrt{\frac{W_N}{N}}} \text{ gilt.}$$

Für $\delta = 0$ ergibt sich insbesondere die übliche t-Verteilung. Die dezentrale t-Verteilung bildet einen formalen Bestandteil im nächsten Satz. Eine analoge Ausführung für Äquivalenztests wird in Wellek (2010), S.129ff. dargestellt. Die Darstellung des nachfolgenden Relevanz-Tests ist an Müller und Denecke (2013), S.223ff angelehnt.

Satz 2.9 (Zweistichproben-Relevanz-t-Test).
Seien $V_1, ..., V_M, W_1, ..., W_N$ unabhängige Zufallsvariablen mit $V_m \sim \mathcal{N}(\mu_1, \sigma^2)$ für $m = 1, ..., M$ und $W_n \sim \mathcal{N}(\mu_2, \sigma^2)$ für $n = 1, ..., N$ mit unbekannten Parametern $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+$ und gegebenem $\delta \geq 0$. Ferner sei

$$T(V, W) = T(V_1, ..., V_M, W_1, ..., W_N) = K \frac{\overline{V.} - \overline{W.}}{\widehat{\sigma}_{V,W}}$$

die Teststatistik. Dann liefert die folgende Entscheidungsregel

$$\varphi(v, w) = \varphi(v_1, ..., v_M, w_1, ..., w_N) = \mathbb{1}_{\{|T(v,w)|>c_\alpha\}} \text{ zum Hypothesenpaar}$$
$$H_0 : |\mu_1 - \mu_2| \leq \delta\sigma \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > \delta\sigma$$

ein Test zum Niveau α, wobei $c_\alpha \in \mathbb{R}$ folgende Gleichung erfüllen soll

$$\alpha = 1 - F(c_\alpha) + F(-c_\alpha)$$

und F die Verteilungsfunktion der dezentralen t-Verteilung mit $M+N-2$ Freiheits-

graden und Nichtzentralitätsparameter $K\delta$ ist.

Beweis. Sei $\mu = (\mu_1, \mu_2)$ ein Parameter aus dem Annahmebereich, das heißt $|\mu_1 - \mu_2| \leq \delta\sigma$. Zu zeigen ist, dass der Fehler erster Art durch das Signifikanzniveau α beschränkt wird. Der Beweis wird in zwei Schritte gegliedert:

1. Man zeige, dass die Gütefunktion $E_\mu(\varphi)$ bezüglich $\mu_1 - \mu_2$ ein monoton wachsendes bzw. fallendes Verhalten für $\mu_1 - \mu_2 \geq 0$ bzw. $\mu_1 - \mu_2 \leq 0$ aufweist.

2. Mit Schritt 1 genügt es dann zu zeigen, dass der Fehler erster Art auf dem Rand des Annahmebereichs, also Parameter mit $|\mu_1 - \mu_2| = \delta\sigma$, genau das Signifikanzniveau α ist.

Schritt 1: Zunächst werden folgende Notationen eingeführt:

$$M := M(V, W) := K(\overline{V.} - \overline{W.}) \text{ und}$$
$$S := S(V, W) := \widehat{\sigma}_{V,W}.$$

Dann lässt sich die Teststatistik durch $T(V, W) = \frac{M}{S}$ darstellen. Nun ist zu zeigen, dass für den Aspekt $a(\mu) := \mu_1 - \mu_2$ mit $a(\mu) \geq 0$ die Gütefunktion $E_\mu(\varphi)$ in dem Sinne monoton wachsend ist, dass für $a(\mu) \leq a(\mu^*)$ folgendes Wachstumsverhalten $E_\mu(\varphi) \leq E_{\mu^*}(\varphi)$ für entsprechende μ, μ^* aus dem Annahmebereich vorliegt (Fall 1). Analog soll für $a(\mu) := \mu_1 - \mu_2 \leq 0$ mit $a(\mu) \leq a(\mu^*) \leq 0$ gelten, dass ein abfallendes Verhalten mit $E_\mu(\varphi) \geq E_{\mu^*}(\varphi)$ vorliegt (Fall 2).
Zum Beweis beider Fälle wird als Vorbereitung die Gütefunktion in eine andere Darstellung überführt:

$$\begin{aligned} E_\mu(\varphi) &= P_\mu(|T(V, W)| > c_\alpha) \\ &= P_\mu\left(\left|\frac{M}{S}\right| > c_\alpha\right) \\ &= \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}\left\{\left|\frac{m}{s}\right| > c_\alpha\right\} \, dP_{\mu,(M,S)}(m, s), \end{aligned}$$

wobei $P_{\mu,(M,S)}$ die gemeinsame Verteilung von M, S bezüglich dem Parameter μ sei. Die gemeinsame Verteilung lässt sich durch $P_{\mu,(M,S)} = P_{\mu,M|S} \otimes P_{\mu,S}$ als bedingtes

Produktmaß schreiben (Klenke (2006), S.178). Damit lässt sich die obige Rechnung fortsetzen

$$
\begin{aligned}
&= \int_{\mathbb{R}\times\mathbb{R}_+} \mathbb{1}\left\{\left|\frac{m}{s}\right| > c_\alpha\right\} d(P_{\mu,M|S=s} \otimes P_{\mu,S})(m,s) \\
&= \int_{\mathbb{R}_+}\int_{\mathbb{R}} \mathbb{1}\left\{\left|\frac{m}{s}\right| > c_\alpha\right\} dP_{\mu,M|S=s}(m) dP_{\mu,S}(s) \\
&= \int_{\mathbb{R}_+} P_\mu\left(\left|\frac{M}{S}\right| > c_\alpha \mid S = s\right) dP_{\mu,S}(s).
\end{aligned}
$$

Man beachte die Verwendung des Satzes von Fubini (Klenke (2006), S.265), um zur zweiten Zeile zu gelangen. Ferner gilt nach Lemma 2.2, dass M und S stochastisch unabhängig sind. Daher lässt sich die Bedingung in der bedingten Wahrscheinlichkeit im Integrand einsetzen, da sie unabhängig von M ist:

$$
\begin{aligned}
&= \int_{\mathbb{R}_+} P_\mu\left(\left|\frac{M}{s}\right| > c_\alpha\right) dP_{\mu,S}(s) \\
&= \int_{\mathbb{R}_+} \left(1 - P_\mu\left(\frac{M}{s} \leq c_\alpha\right) + P_\mu\left(\frac{M}{s} \leq -c_\alpha\right)\right) dP_{\mu,S}(s).
\end{aligned}
$$

Weiterhin gilt $\frac{1}{\sigma}(M - a(\mu)) \sim \mathcal{N}(0,1)$, wenn μ der wahre Parameter ist. Die obigen Ereignisse können durch $P_\mu(\frac{1}{\sigma}(M - a(\mu)) \leq \frac{1}{\sigma}(sc_\alpha - a(\mu)))$ bzw. $P_\mu(\frac{1}{\sigma}(M - a(\mu)) \leq \frac{1}{\sigma}(-sc_\alpha - a(\mu)))$ äquivalent dargestellt werden, sodass die Zufallsvariablen einer Standardnormalverteilung folgen:

$$
= \int_{\mathbb{R}_+} \underbrace{\left(1 - \Phi\left(\frac{1}{\sigma}\left(sc_\alpha - a(\mu)\right)\right) + \Phi\left(\frac{1}{\sigma}\left(-sc_\alpha - a(\mu)\right)\right)\right)}_{=:g_{\sigma,s}(a(\mu))} dP_{\mu,S}(s), \qquad \text{(I)}
$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung sei. Für den ersten Fall muss nun gezeigt werden, dass für $a(\mu) \leq a(\mu^*)$ die Funktion $g_{\sigma,s}(a(\mu))$ monoton wachsend ist. Dann würde mit (I) folgen, dass $E_\mu(\varphi) \leq E_{\mu^*}(\varphi)$.
Sei f die Dichtefunktion der Stanardnormalverteilung. Die Monotonie von $g_{\sigma,s}$ wird

mithilfe ihrer Ableitung gezeigt:

$$\begin{aligned}
\frac{\mathrm{d}g_{\sigma,s}(a(\mu))}{\mathrm{d}a(\mu)} &= \frac{1}{\sigma}\left(\Phi'\left(\frac{1}{\sigma}\left(sc_\alpha - a(\mu)\right)\right) - \Phi'\left(\frac{1}{\sigma}\left(-sc_\alpha - a(\mu)\right)\right)\right) \\
&= \frac{1}{\sigma}\left(f\left(\frac{1}{\sigma}\left(sc_\alpha - a(\mu)\right)\right) - f\left(\frac{1}{\sigma}\left(-sc_\alpha - a(\mu)\right)\right)\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\left(\exp\left(-\frac{(sc_\alpha - a(\mu))^2}{2\sigma^2}\right) - \exp\left(-\frac{(-sc_\alpha - a(\mu))^2}{2\sigma^2}\right)\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\left(\exp\left(-\frac{(sc_\alpha - a(\mu))^2}{2\sigma^2}\right) - \exp\left(-\frac{(sc_\alpha + a(\mu))^2}{2\sigma^2}\right)\right) \overset{!}{\geq} 0.
\end{aligned}$$

Die Behauptung, dass letzterer Ausdruck größer 0 ist, lässt sich elementar durch Äquivalenzumformungen nachprüfen. Man beachte dazu, dass $\sigma, s, c_\alpha, a(\mu) \geq 0$ sind:

$$\begin{aligned}
& \exp\left(-\frac{(sc_\alpha - a(\mu))^2}{2\sigma^2}\right) \overset{!}{\geq} \exp\left(-\frac{(sc_\alpha + a(\mu))^2}{2\sigma^2}\right) \\
\Leftrightarrow\ & (sc_\alpha - a(\mu))^2 \overset{!}{\leq} (sc_\alpha + a(\mu))^2 \\
\Leftrightarrow\ & (sc_\alpha)^2 - 2sc_\alpha a(\mu) + a(\mu)^2 \overset{!}{\leq} (sc_\alpha)^2 + 2sc_\alpha a(\mu) + a(\mu)^2 \\
\Leftrightarrow\ & 0 \overset{!}{\leq} 4sc_\alpha a(\mu),
\end{aligned}$$

wobei letzte Aussage wegen der Bedingungen an die Parameter wahr ist und gefolgert werden kann, dass $g_{\sigma,s}$ monoton wachsend für positive $a(\mu)$ ist.
Der zweite Fall kann analog für das umgekehrte Relationszeichen nachgerechnet werden. Es ergibt sich $0 \overset{!}{\geq} 4sc_\alpha a(\mu)$, was eine wahre Aussage ist, da $a(\mu) \leq 0$ ist. Daraus folgt, dass $g_{\sigma,s}$ für negative $a(\mu)$ monoton fallend ist. Aus (I) folgt dann: $E_\mu(\varphi) \geq E_{\mu^*}(\varphi)$, falls $a(\mu) \leq a(\mu^*) \leq 0$.

Schritt 2: Nun wird gezeigt, dass auf dem Rand der Nullhypothese $|\mu_1 - \mu_2| = \delta\sigma$ das Signifikanzniveau α exakt eingehalten wird. Sei $a(\mu) = \mu_1 - \mu_2 = \delta\sigma$. Der Fall

$\mu_1 - \mu_2 = -\delta\sigma$ kann genauso bewiesen werden. Es gilt:

$$\begin{aligned} P_\mu\left(\left|K\frac{\overline{V.}-\overline{W.}}{\widehat{\sigma}_{V,W}}\right| > c_\alpha\right) &= P_\mu\left(\left|K\frac{\overline{V.}-\overline{W.}-a(\mu)+a(\mu)}{\widehat{\sigma}_{V,W}}\right| > c_\alpha\right) \\ &= P_\mu\left(\left|K\frac{\overline{V.}-\overline{W.}-a(\mu)+\delta\sigma}{\widehat{\sigma}_{V,W}}\right| > c_\alpha\right) \\ &= P_\mu\left(\left|\frac{\frac{K}{\sigma}(\overline{V.}-\overline{W.}-a(\mu))+K\delta}{\frac{\widehat{\sigma}_{V,W}}{\sigma}}\right| > c_\alpha\right). \end{aligned}$$

Da $\frac{K}{\sigma}(\overline{V.}-\overline{W.}-a(\mu)) \sim \mathcal{N}(0,1)$ ist, folgt unter Verwendung von Lemma 2.2, dass der linke Ausdruck (ohne Beträge) einer dezentralen t-Verteilung mit Nichtzentralitätsparameter $K\delta$ folgt. Schließlich lässt sich die Wahrscheinlichkeit wie folgt umformen:

$$\begin{aligned} &= 1 - P_\mu\left(\frac{\frac{K}{\sigma}(\overline{V.}-\overline{W.}-a(\mu))+K\delta}{\frac{\widehat{\sigma}_{V,W}}{\sigma}} \leq c_\alpha\right) \\ &\quad + P_\mu\left(\frac{\frac{K}{\sigma}(\overline{V.}-\overline{W.}-a(\mu))+K\delta}{\frac{\widehat{\sigma}_{V,W}}{\sigma}} \leq -c_\alpha\right) = \alpha, \end{aligned}$$

wobei die letzte Gleichheit nach Definition von c_α gilt. Wegen der nachgewiesenen Monotonieeigenschaft der Gütefunktion aus Schritt 1 ist der Fehler erster Art für beliebige $a(\mu)$ mit $0 \leq a(\mu) \leq \delta\sigma$ durch α beschränkt. Analog nutzt man die Monotonieeigenschaft, um nachzuweisen, dass der Fehler erster Art für $a(\mu)$ mit $-\delta\sigma \leq a(\mu) \leq 0$ durch α beschränkt, womit gezeigt ist, dass der Test das Signifikanzniveau α einhält. □

Es werden nun einige Bemerkungen zur Anwendung dieses t-Tests geliefert.

Konsequenzen bei falscher Wahl der Varianz

Im Vergleich zu den anderen beiden t-Tests ist für einen festen Relevanzparameter $\delta \geq 0$ das Hypothesenpaar zusätzlich von der unbekannten Standardabweichung σ abhängig. Hat man keine Informationen über die Standardabweichungen, weiß man bei der Anwendung dieses Tests nicht, welches Hypothesenpaar man tatsächlich untersucht, da man ein zur Standardabweichung passenden Relevanzparameter δ wählen muss. Es kann problematisch sein, wenn man die Standardabweichung zu stark unter- oder überschätzt, da so δ auch verzerrt wird.

Unterschätzt man die Varianz, so ist der eigentlich untersuchte Nicht-Relevanzbereich $[-\delta, \delta]$ größer. Der Test wird dann zum vermeintlichen Hypothesenpaar seltener ablehnen, da ein größerer Nicht-Relevanzbereich als nötig überschritten werden muss. Überschätzt man die Varianz, passiert das Gegenteil. Der wahre Nicht-Relevanzbereich ist kleiner, wodurch man häufiger zum vermeintlichen Hypothesenpaar ablehnt. Vor allem letzteres führt dann im Allgemeinen zu einem Nichteinhalten des Niveaus unter dem vermeintlichen Hypothesenpaar.

Berechnung von c_α

Der Wert $c_\alpha = c_\alpha(M, N, \delta, \alpha)$ muss zur Anwendung des Tests bestimmt werden. Er hängt von der Summe der Stichprobenumfänge $M + N$, dem Relevanzparameter δ und dem Signifikanzniveau α ab. Nach Definition ist c_α eine reelle Zahl, sodass

$$\begin{aligned} \alpha &\stackrel{!}{=} 1 - F(c_\alpha) + F(-c_\alpha) \\ \Leftrightarrow \quad 0 &\stackrel{!}{=} 1 - F(c_\alpha) + F(-c_\alpha) - \alpha \end{aligned}$$

gilt. Dabei ist F die Verteilungsfunktion der dezentralen t-Verteilung mit $M+N-2$ Freiheitsgraden und Nichtzentralitätsparameter $K\delta$. c_α ist eindeutig, da die Funktion $g(c) := 1 - F(c) + F(-c) - \alpha$ streng monoton fallend ist. Das sieht man ein, indem man die Ableitung $g'(c) = -f(c) - f(-c)$ betrachtet, wobei f die Dichtefunktion zur Verteilungsfunktion von F ist. Da die Dichtefunktion strikt positiv ist, also $f(c) > 0$ für jedes $c \in \mathbb{R}$ gilt, ist $g'(c) < 0$ für jedes $c \in \mathbb{R}$. Ferner lässt sich c_α somit als folgendes Minimierungsproblem mit eindeutiger Lösung schreiben:

$$c_\alpha := \operatorname*{argmin}_{c \in \mathbb{R}} |1 - F(c) + F(-c) - \alpha|.$$

Die Existenz der Lösung ist gewährleistet, da $\alpha \in (0, 1)$ und die stetige Funktion $1 - F(c) + F(-c)$ Werte zwischen $(0, 2)$ für $c \in \mathbb{R}$ annimmt. Mit der `R`-Funktion `pt()` kann die Verteilungsfunktion der dezentralen t-Verteilung über `ncp` zur Angabe des Dezentralitätsparameters berechnet werden. Der kritische Wert c_α kann dann über eine Optimierung bestimmt werden.

Die Tabelle 1 listet einige Werten für c_α bei $\delta = 1$ und $\alpha = 0.05$ und verschiedene

Stichprobenumfänge M, N, die in Kapitel 3 der Arbeit verwendet werden.

Tabelle 1: Einige c_α bei $\delta = 1$, $\alpha = 0.05$ und Gesamtstichprobenumfang $M + N$

Stichprobenumfang von M (wobei: $M = N$)	5	10	20	30	50	100	200
$c_{0.05}$	3.962	4.303	5.094	5.759	6.846	8.882	11.789

Zum Abschluss soll der Fall $\delta = 0$ diskutiert werden. Auch hier wird sich der übliche Zweistichproben-t-Test ergeben. Man betrachte das Minimierungsproblem:

$$c_\alpha = \underset{c \in \mathbb{R}}{\operatorname{argmin}} \, |1 - F(c) + F(-c) - \alpha|$$

Da $K\delta = 0$ ist, liefert F die Verteilungsfunktion der zentralen t-Verteilung mit $M + N - 2$ Freiheitsgrad. Die Symmetrie der zentralen t-Verteilung liefert:

$$\begin{aligned} c_\alpha &= \underset{c \in \mathbb{R}}{\operatorname{argmin}} \, |1 - F(c) + (1 - F(c)) - \alpha| \\ &= \underset{c \in \mathbb{R}}{\operatorname{argmin}} \, |2 - 2F(c) - \alpha| \\ &= \underset{c \in \mathbb{R}}{\operatorname{argmin}} \left|1 - F(c) - \frac{\alpha}{2}\right| \\ &= t_{M+N-2;1-\frac{\alpha}{2}}, \end{aligned}$$

wobei $t_{M+N-2;1-\frac{\alpha}{2}}$ das $(1-\frac{\alpha}{2})$-Quantil der t-Verteilung mit $M+N-2$ Freiheitsgraden ist. Setzt man $c_\alpha = t_{M+N-2;1-\frac{\alpha}{2}}$ in die Entscheidungsregel aus Satz 2.9 ein, liefert das gerade den Zweistichproben-t-Test aus Satz 2.4(a).

2.3 Zweistichproben-Relevanz-Tests basierend auf Datentiefen

Die Einführung der Relevanz-Tests basierend auf Datentiefen soll durch die Abbildung 1 motiviert werden. In der Darstellung liegen zwei Lageparameter $\mu = (\mu_1, \mu_2)$ vor, um die unabhängig, identische Realisationen von Zufallsvariabeln streuen (in der Abbildung 1 mit Varianz $\sigma^2 = \frac{1}{2}$ unter einer Normalverteilung). Dabei sei zu beachten, dass die Parameter μ_1, μ_2 unbekannt sind, sodass aus den Realisationen auf die Lageparameter geschlossen werden muss.

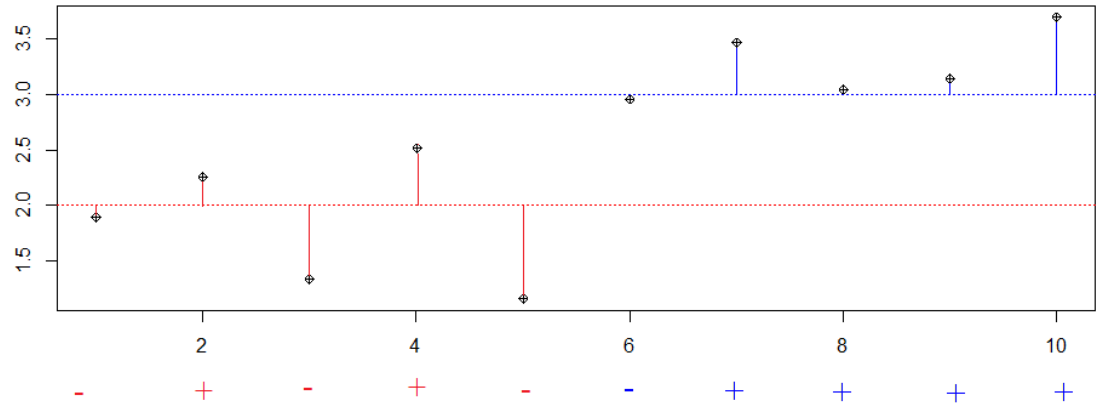


Abbildung 1: Fünf Daten aus jeweils zwei Stichproben mit Mittelwert $\mu_1 = 1$ und $\mu_2 = 2$ aus einer Normalverteilung mit Abstand $\delta = 1$ mit Varianz $\sigma^2 = \frac{1}{2}$ mit eingezeichneten Residuen und ihren Vorzeichen

Man betrachtet nun ein fixes Parameterpaar μ, das im Nicht-Relevanzbereich zum Relevanzparameter δ liegt, also für $\delta \geq 0$ gelte $|\mu_1 - \mu_2| \leq \delta$. Das Verhalten der Residuen für das ausgewählte Parameterpaar im Nicht-Relevanzbereich wird im Folgenden untersucht. Die Residuen seien dabei wie folgt definiert.

$$\text{Für } z_n := \begin{cases} x_n & \text{für } n = 1, ..., M \\ y_{n-M} & \text{für } n = M+1, ..., N+M \end{cases} \quad \text{sei}$$

$$\text{res}(\mu, z_n) := \begin{cases} x_n - \mu_1 & \text{für } n = 1, ..., M \\ y_{n-M} - \mu_2 & \text{für } n = M+1, ..., N+M \end{cases} \quad \text{für } \mu = (\mu_1, \mu_2) \in \mathbb{R}^2$$

Aus der Gestalt der Residuen können Informationen gewonnen werden, inwieweit die gewählten Lageparamtern den wirklichen entsprechen. Beispielsweise betrachtet man quadrierte Summen der Residuen oder ihre Vorzeichen. Unter der Annahme, dass die Fehler $E_1, ..., E_{M+N}$ den Median 0 haben, erwartet man, dass für die wahren Parameter ungefähr gleich viele positive und negative Vorzeichen der Residuen auftreten. Dadurch könnte man einen Vorzeichentest motivieren:
Man untersucht das Verhalten der Vorzeichen der Residuen für alle Parameterpaare μ im Nicht-Relevanzbereich. Liegen *für alle Parameter im Nicht-Relevanzbereich* untypische Verhältnisse der Vorzeichen vor, so spricht das dafür, dass die Parameter μ_1 und μ_2 einen relevanten Unterschied besitzen.

Dieser Ansatz weist allerdings das Problem auf, dass man beliebige Parameterpaare μ_1, μ_2 im Nicht-Relevanzbereich betrachten muss. Dadurch findet man immer ein Paar im Annahmebereich, sodass die Nullhypothese nicht verworfen werden kann. Weitere Details lassen sich in Abschnitt 2.3.4 zu diesem Ansatz nachlesen.
Eine andere Idee ist die Betrachtung von *Vorzeichenwechsel* der Residuen. Ziel ist es nun ein Parameterpaar μ im Nicht-Relevanzbereich zu finden, sodass möglichst viele Vorzeichenwechsel generiert werden. Viele Vorzeichenwechsel sprechen für eine gute Anpassung des Lageparameters. Betrachtet man in Abbildung 1 das wahre Parameterpaar und $\delta = 1$ als Relevanzparameter, so liegen zum Beispiel relativ viele Vorzeichenwechsel vor. Die Annahme, $E_1, ..., E_{M+N}$ haben den Median 0, liefert die Erwartungshaltung, dass die Hälfte aller benachbarten Datenpaare unter den wahren Parametern ihre Vorzeichen wechseln sollten. Fünf Vorzeichenwechsel entsprechen demnach einer sehr guten Anpassung. Die Annahme, dass der Parameter $\mu = (1, 2)$ im Nicht-Relevanzbereich liegen, würde man nicht ablehnen.

Wählt man mit $\delta \leq \frac{1}{2}$ kleiner, so lassen sich, wie in Abbildung 2 zu sehen, weni-

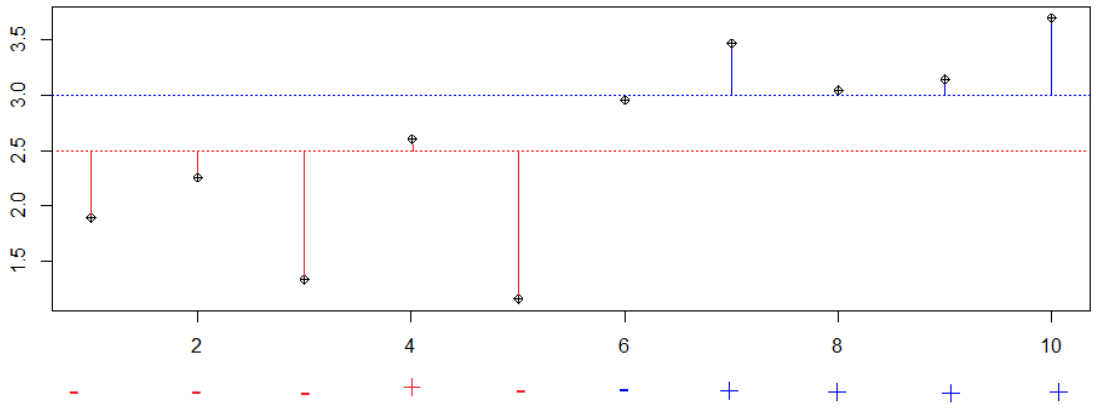


Abbildung 2: Fünf Daten aus jeweils zwei Stichproben mit Mittelwert $\mu_1 = 1$ und $\mu_2 = 2$ aus einer Normalverteilung mit Abstand $\delta = \frac{1}{2}$ mit Varianz $\sigma^2 = \frac{1}{2}$ mit eingezeichneten Residuen und ihren Vorzeichen

ger Möglichkeiten an Vorzeichenwechsel generieren, da die gewählten Lageparameter näher aneinander liegen müssen. Damit kann eventuell eine gute Anpassung der Lageparameter nicht gewährleistet werden. Existiert kein einziges Parameterpaar, das viele Vorzeichenwechsel erzeugen kann, würde man die Annahme, dass die wahren Parameter im Nicht-Relevanzbereich liegen, ablehnen.

Die Schwierigkeit bei diesem Ansatz ist der Fall, wenn man die Nullhypothese ablehnen soll, da eine Nichtexistenz deutlich schwieriger nachzuweisen ist als eine Existenz. Für eine Nichtexistenz müssen alle infrage kommende Parameter im Nicht-Relevanzbereich überprüft werden. Diese Hürde erlaubt aber gerade, im Vergleich zum oben genannten Idee eines Vorzeichentests, einen sinnvollen Tests bezüglich der Vorzeichenwechsel zu konstruieren.

2.3.1 Die volle Dreiertiefe

Die Untersuchung der Vorzeichenwechsel kann durch verschiedene Größen von Tupeln erfolgen. Man kann beispielsweise Paare, Tripel usw. von Vorzeichen der Residuen betrachten. Betrachtet man die Vorzeichen aller möglichen unterschiedlichen Residuen bis auf ihre Reihenfolge, so spricht man von einer *vollen Tiefe*. Im Folgenden wird für die Fehler $E_1, ..., E_{M+N}$ zusätzlich eine stetige Verteilung angenommen, sodass fast sicher kein Residuum gleich 0 ist. Unter dieser Annahme kann folgende Definition eingeführt werden (Kustosz, Müller und Wendler (2016)).

Definition 2.10 (Volle Daten-Tiefe).
Gegeben seien Realisationen zweier Stichproben $x_1, ..., x_M, y_1, ..., y_N$, auch ausgedrückt durch $z = (z_1, ..., z_{M+N})$ wie oben. Für ein gegebenes $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$ werden folgende Maßzahlen eingeführt:
(a) Die volle $(K+1)$-Tiefe für $K \in \mathbb{N}$ hat die Gestalt:

$$d_{M+N;(K)}(\mu, z) = \frac{1}{\binom{M+N}{K+1}} \sum_{1 \le n_1 < n_2 < ... < n_{K+1} \le M+N} \left(\prod_{k=1}^{K+1} \mathbb{1}\{res(\mu, z_{n_k})(-1)^k > 0\} + \prod_{k=1}^{K+1} \mathbb{1}\{res(\mu, z_{n_k})(-1)^{k+1} > 0\} \right).$$

(b) Für den Fall $K = 2$ ergibt sich die volle Dreier-Tiefe:

$$d_{M+N}(\mu, z) = \frac{1}{\binom{M+N}{3}} \sum_{1 \le n_1 < n_2 < n_3 \le M+N} \left(\prod_{k=1}^{3} \mathbb{1}\{res(\mu, z_{n_k})(-1)^k > 0\} + \prod_{k=1}^{3} \mathbb{1}\{res(\mu, z_{n_k})(-1)^{k+1} > 0\} \right).$$

Der Fall, dass Residuen gleich 0 sind, braucht nicht betrachtet werden, da dies fast sicher wegen der stetigen Verteilung von $E_1, ..., E_{M+N}$ nicht auftritt. Streng betrachtet reicht es zu fordern, dass die $P(\text{res}(\mu, Z_n) = 0) = 0$ für beliebige μ, z_n gelten soll, wodurch die Anforderungen an $E_1, ..., E_{M+N}$ abgeschwächt werden könnten (und beispielsweise $E_1, ..., E_{M+N}$ auch diskret sein dürfen).
Die volle $(K+1)$-Tiefe entspricht der relativen Anzahl aller geordneten Tupeln von Residuen der Größe $K+1$, bei denen K Vorzeichenwechsel im Tupel vorliegen, in Relation zu allen möglichen Tupeln bis auf Reihenfolge. Eine besondere Rolle spielt die volle Dreier-Tiefe, da ihr asymptotisches Verhalten bekannt ist und somit zur Konstruktion von Testverfahren genutzt werden kann.

Asymptotik der vollen Dreier-Tiefe und Konstruktion des Testverfahrens

Die Asymptotik der normierten vollen Dreier-Tiefe wird in Kustosz, Müller und Leucht (2016) für das AR(1)-Modell hergeleitet. Es kann aber auch auf andere Modellansätze, wie das hier betrachtete Zweistichproben-Modell für zwei verschiedene Lageparameter, verwendet werden, da bei der Herleitung der Asymptotik lediglich die Eigenschaften der Residuen berücksichtigt werden. In der zitierten Quelle wird für die normierte volle Dreier-Tiefe $T_{M+N}(\mu, Z)$ folgende Konvergenz in Verteilung hergeleitet:

$$\begin{aligned} T_{M+N}(\mu, Z) :=& (M+N)\left(d_{M+N}(\mu, Z) - \frac{1}{4}\right) \\ &\xrightarrow[(M+N)\to\infty]{\mathcal{D}} \frac{3}{4} + \frac{3}{4}G_2(0)^2 - \frac{3}{2}\int_{-2}^{-2} G_1(t)^2\, dt. \end{aligned}$$

Bei $(G(t))_{t\in[-2,2]} = (G_1(t), G_2(t))_{t\in[-2,2]}$ handelt es sich um einen zweidimensionalen, zentrierten Gauß-Prozess mit folgender Kovarianzstruktur für $s, t \in [-2, 2]$

$$\text{Cov}(G(t), G(s)) = \begin{pmatrix} \int_0^1 \mathbb{1}_{(-0.5,0.5]}(x-s)\mathbb{1}_{(-0.5,0.5]}(x-t)\, dx & \int_0^1 \mathbb{1}_{(-0.5,0.5]}(x-s)\, dx \\ \int_0^1 \mathbb{1}_{(-0.5,0.5]}(x-t)\, dx & 1 \end{pmatrix}.$$

Man beachte, dass die asymptotische Verteilung nicht von einem Zeitpunkt $t \in [-2, 2]$ abhängt, da über diesen Zeitbereich integriert wird und daher eine reellwertige Zufallsgröße bildet. Zur Konstruktion eines asymptotischen Tests können somit

die Quantile q_α der asymptotischen Verteilung der normierten vollen Dreier-Tiefe bestimmt werden. Die Quantile lassen sich in `R` mit dem Paket `rexpar` von Kustosz und Szugat (2016) mit dem Befehl `SimQuants()` numerisch berechnen, um sie für das Testverfahren in Satz 2.11 zu nutzen. Einige Quantile von praxisrelevanten Signifikanzniveaus werden in Tabelle 2 aufgeführt.

Tabelle 2: Quantile der asymptotischen Verteilung für verschiedene Signifikanzniveaus α

Signifikanzniveau α	0.05	0.01	0.001
Quantil q_α	-1.255	-2.240	-3.714

Satz 2.11 (Zweistichproben-Relevanz-Test mit voller Dreier-Tiefe).
Seien $X_1, ..., X_M$ und $Y_1, ..., Y_N$ beliebige, stetige Zufallsvariablen, die den Generalannahmen aus Abschnitt 2 genügen. Man beachte die Notation $Z = (X, Y)$ für die zusammengefassten Stichproben. Sei das Hypothesenpaar $H_0 : |\mu_1 - \mu_2| \leq \delta$ gegen $H_1 : |\mu_1 - \mu_2| > \delta$ mit Annahmebereich $\Theta_0 := \{\mu \in \mathbb{R}^2 \mid |\mu_1 - \mu_2| \leq \delta\}$ gegeben. Folgende Entscheidungsregel liefert dann einen asymptotischen Test zum Niveau α:

$$\textit{Man verwerfe } H_0, \textit{ falls } \sup_{\mu \in \Theta_0} T_{M+N}(\mu, Z) < q_\alpha.$$

Beweis. Man betrachte ein beliebiges $\mu^* \in \Theta_0$. Dann folgt

$$P_{\mu^*}(\sup_{\mu \in \Theta_0} T_{M+N}(\mu, Z) < q_\alpha) \leq P_{\mu^*}(T_{M+N}(\mu^*, Z) < q_\alpha) \xrightarrow[(M+N)\to\infty]{} \alpha$$

durch Einsetzen von μ^* als Kandidaten für das Supremum. □

An dieser Stelle ist zu betonen, dass die Annahme $\text{med}(E_i) = 0$ für $i = 1, ..., M+N$ entscheidend für die Asymptotik der Teststatistik ist.
Weiterhin sei zu erwähnen, dass für den Fall der vollen Zweier-Tiefe mit $K = 1$ die asymptotische Verteilung in Kustosz und Müller (2014) hergeleitet wurde und der verschobenen und skalierten χ^2-Verteilung $\frac{1}{2}(1 - U)$ mit $U \sim \chi^2_1$ entspricht. Die Herleitung wird in der angegebenen Quelle für das AR(1)-Modell durchgeführt; allerdings ist die Rechnung auch auf beliebige Modelle, die den Generalannahmen genügen und fast sicher Residuen ungleich 0 besitzen, übertragbar. Ferner können die

Quantile der normierten vollen Tiefe auch für $K \geq 4$ durch Simulationen berechnet und weitere Testverfahren gebildet werden.

Geschwindigkeit der Asymptotik der vollen Dreier-Tiefe

Zum Abschluss soll die Asymptotik des Tests genauer untersucht werden. Bei zu kleinen Stichprobenumfängen bildet das Testverfahren keinen sinnvollen Test, da er nicht ablehnen kann. Dies hängt damit zusammen, dass bei zu kleinen Stichprobenumfängen keine Realisation der Teststatistik den kritischen Wert q_α überhaupt unterschreiten kann. Ziel ist es also den niedrigsten Stichprobenumfang zu finden, bei dem der Test ablehnen kann. Man betrachte die Teststatistik

$$(M+N)\left(d_{M+N}(\mu, z) - \frac{1}{4}\right).$$

Für einen festen Stichprobenumfang wird der minimal annehmbare Wert der Teststatistik mit q_α verglichen. Die Teststatistik wird genau dann minimal, wenn die volle Dreier-Tiefe $d_{M+N}(\mu, z) = 0$ ist. Folgende notwendige Bedingung muss also erfüllt werden damit das Testverfahren die Nullhypothese ablehnen kann:

$$-\frac{M+N}{4} \overset{!}{\leq} q_\alpha \Rightarrow M+N \overset{!}{\geq} -4q_\alpha$$

Für verschiedene typische Signifikanzniveaus α gibt die nachfolgende Tabelle 3 den minimalen Stichprobenumfang an, sodass das Testverfahren überhaupt ablehnen kann. Die Werte für die Stichprobenumfänge sind dabei auf die nächste natürliche Zahl aufgerundet worden.

Signifikanzniveau α	0.05	0.01	0.001
Stichprobenumfang für $M+N$	6	9	15

Tabelle 3: notwendiger Stichprobenumfang $M+N$ für ein gegebenes Signifikanzniveau α bei der vollen Dreier-Tiefe

Die notwendigen Stichprobenumfänge sind relativ klein. Allerdings beachte man, dass die Asymptotik dennoch durch zu wenige Daten beeinträchtigt werden könnte und hier lediglich Mindestanforderungen hergeleitet wurden. Weitere Details zur

Auswirkung des Stichprobenumfangs auf die Asymptotik und die Qualität des Testverfahrens lassen sich in der Simulationsstudie in Kapitel 3 nachlesen.
Es sei zu betonen, dass die Berechnung der vollen Dreier-Tiefe für hohe Datenmenge sehr hohe Rechenzeiten benötigt, da insgesamt $2\binom{M+N}{3}$ Summanden auf ihre Vorzeichen überprüft werden müssen. Daher wird im Abschnitt 2.3.2 eine alternative, weniger rechenintensive Datentiefe vorgestellt.

Zusammenhang zur Simplex-Tangent-Tiefe

In diesem Unterabschnitt soll ein Zusammenhang zwischen der vollen Datentiefe zur Simplex-Tangent-Tiefe angegeben werden (siehe auch Kustosz, Müller und Wendler (2016)). Man betrachte folgendes Regressionsmodell

$$y_n = g(x_n, \vartheta) + e_n \text{ für } n = 1, ..., N,$$

das durch einen K-dimensionalen Parametervektor $\vartheta \in \mathbb{R}^K$ in Abhängigkeit von den beobachteten Datenpunkten $(x_1, y_1), ..., (x_N, y_N)$ beschrieben wird. Die Fehler $e_1, ..., e_N$ seien Realisationen unabhängig, identisch verteilter Zufallsvariablen $E_1, ..., E_N$. Ferner seien die Regressoren $x_1, ..., x_N$ als Realisationen von Zufallsvariablen $X_1, ..., X_N$ zu interpretieren, sodass

$$X_1 < ... < X_N$$

fast sicher die angegebene Ordnungsstruktur erfüllt ist. Der Begriff der *Tangent-Tiefe*, welcher unter anderem in Rousseeuw und Hubert (1999) und Mizera (2002) eingeführt wird, hat die Idee der Betrachtung von Vorzeichenwechsel motiviert. Für einen Parameter ϑ und $z^* = (z_1, ..., z_{r+1})$ ist sie definiert durch

$$d_T(\vartheta, z^*) := \frac{1}{K+1} \min_{u \in \mathbb{R}^K} \# \left\{ n \in \{1, ..., K+1\} \mid u^\top \frac{\partial}{\partial \vartheta} \text{res}(\vartheta, z_n)^2 \leq 0 \right\},$$

wobei $\text{res}(\vartheta, z_n) := y_n - g(x_n, \vartheta)$ für $n = 1, ..., N$ die Residuen seien. In Kustosz, Müller und Wendler (2016) wird im Theorem 1 gezeigt, dass unter gewissen Regularitäten an die Regressionsfunktion g sich die Bedingung $d_T(\vartheta, z^*) > 0$ äquivalent zu

alternierenden Vorzeichen der Residuen (mit der Zusatzannahme, dass die Residuen fast sicher ungleich 0 sind) charakterisieren lässt.
Die aus der Tangent-Tiefe abgeleitete *Simplex-Tangent-Tiefe* von ϑ in $z^* = (z_1, ..., z_N)$ ist wie folgt definiert:

$$d_{SIM}(\vartheta, z^*) := \frac{1}{\binom{N}{K+1}} \sum_{1 \leq n_1 < n_2 < ... < n_{K+1} \leq N} \mathbb{1}\{d_T(\vartheta, z_{n_1}, ..., z_{n_{K+1}}) > 0\}.$$

Besitzt das Regressionsmodell, die in Kustosz, Müller und Wendler (2016) im Theorem 1 dargestellte Regularität, so liefert das (mit der Zusatzannahme, dass die Residuen fast sicher ungleich 0 sind)

$$d_{SIM}(\vartheta, z^*) = d_{N;(K)}(\mu, z_*),$$

d.h. die Simplex-Tangent-Tiefe und die volle $(K+1)$-Tiefe sind dann identisch. Dieser Zusammenhang gilt für viele Klassen von Regressionsmodellen, wie in Kustosz, Müller und Wendler (2016) gezeigt wird. Der hier dargestellte Zusammenhang motiviert die Betrachtung von Vorzeichenwechsel zur Untersuchung von Regressionstiefen, deren Ursprünge in der globalen Regressionstiefe und der Tangent-Tiefe liegen. Insbesondere kann die Simplex-Tangent-Tiefe nun einfacher über die Vorzeichenwechsel berechnet werden. Sie kann aber auch innermathematisch bei Untersuchungen von asymptotischen Verhalten aus einer anderen Perspektive untersucht werden.

2.3.2 Die vereinfachte Dreier-Tiefe

Statt alle Tupel zu untersuchen, werden bei der vereinfachten Tiefe lediglich benachbarte betrachtet:

Definition 2.12 (Vereinfachte Tiefe).
Gegeben seien Realisationen zweier Stichproben $x_1, ..., x_M, y_1, ..., y_N$, auch ausgedrückt durch $z = (z_1, ..., z_{M+N})$ wie oben. Für ein gegebenes $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$ werden folgende Maßzahlen eingeführt: (siehe auch Kustosz, Müller und Wendler (2016))

(a) Die vereinfachte $(K+1)$-Tiefe für $K \in \mathbb{N}$ hat die Gestalt:

$$d^S_{M+N;(K)}(\mu, z) := \frac{1}{M+N-K} \sum_{n=1}^{M+N-K} \left(\prod_{k=1}^{K+1} \mathbb{1}\{res(\mu, z_{n+k-1})(-1)^k > 0\} + \prod_{k=1}^{K+1} \mathbb{1}\{res(\mu, z_{n+k-1})(-1)^{k+1} > 0\} \right).$$

(b) Für den Fall $K = 2$ ergibt sich die vereinfachte Dreier-Tiefe:

$$d^S_{M+N}(\mu, z) := \frac{1}{M+N-2} \sum_{n=1}^{M+N-2} \left(\prod_{k=1}^{3} \mathbb{1}\{res(\mu, z_{n+k-1})(-1)^k > 0\} + \prod_{k=1}^{3} \mathbb{1}\{res(\mu, z_{n+k-1})(-1)^{k+1} > 0\} \right).$$

Im Gegensatz zur vollen Dreier-Tiefe benötigt die vereinfachte Dreier-Tiefe lediglich $2(M+N-2)$ Überprüfungen und besitzt damit eine niedrigere Rechenzeit.

Asymptotik der vereinfachten Dreier-Tiefe und Konstruktion des Testverfahrens

Das asymptotische Verhalten der vereinfachten Tiefe lässt sich mit einfacheren Methoden nachrechnen als das der vollen Tiefe. Es gibt zwar Abhängigkeiten unter den Summanden bei der Berechnung der vereinfachten Tiefe; jedoch bestehen zwischen früheren und späteren Summanden gar keine Abhängigkeiten mehr, da nur die Vorzeichenwechsel benachbarter Residuen betrachtet werden. Folgende Definition formalisiert diese schwächere Form von stochastischer Abhängigkeit.

Definition 2.13 (m-Abhängigkeit von Zufallsvariablen).
Eine Folge von Zufallsvariablen $(V_n)_{n\in\mathbb{N}}$ heißt m-abhängig für ein $m \in \mathbb{N}_0$, falls $(V_i, ..., V_{i+j})$ stochastisch unabhängig von $(V_{i+j+k}, ..., V_{i+k+j+l})$ für alle $k > m$ für $i, j, l \in \mathbb{N}$ ist.

Für m-abhängige Zufallsvariablen gibt es folgende Konvergenzaussage in Form eines Zentralen Grenzwertsatzes (siehe Hoeffding und Robbins (1948)).

Satz 2.14 (Zentraler Grenzwertsatz für m-abhängige Zufallsvariablen).
Sei $(X_n)_{n\in\mathbb{N}}$ eine Folge m-abhängiger Zufallsvariablen mit

(a) $E(|X_n|^3) \leq R < \infty$ *für alle* $n \in \mathbb{N}$

(b) $\lim_{p\to\infty} \frac{1}{p} \sum_{h=1}^{p} A_{i+h} = A$ *existiert gleichmäßig für alle* $i \in \mathbb{N}_0$, *wobei*

$A_i = Var(X_{i+m}) + 2\sum_{j=1}^{m} Cov(X_{i+m-j}, X_{i+m})$ *für* $i \in \mathbb{N}$ *sei. Dann gilt:*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (X_i - E(X_i)) \xrightarrow[N\to\infty]{} \mathcal{N}(0, A).$$

Unter der Annahme, dass eine Folge m-abhängiger Zufallsvariablen identisch verteilt ist, wird die Voraussetzung (b) des Satzes 2.14 erfüllt, da alle Summanden A_{h+p} gleich sind. Dies wird in der Rechnung im nächsten Beweis verwendet.

Satz 2.15 (Asymptotische Normalität der vereinfachten Tiefe).
Seien $X_1, ..., X_M$ *und* $Y_1, ..., Y_N$ *beliebige, stetige Zufallsvariablen, die den Generalannahmen aus Abschnitt 2 genügen. Man beachte die Notation* $Z = (X, Y)$ *für die zusammengefassten Stichproben. Dann gilt:*

$$T^S_{M+N;(K)}(\mu, Z) = \sqrt{M+N-K} \frac{d^S_{M+N;(K)}(\mu, Z) - \left(\frac{1}{2}\right)^K}{\sqrt{\left(\frac{1}{2}\right)^K \left(3 - 3\left(\frac{1}{2}\right)^K - K\left(\frac{1}{2}\right)^{K-1}\right)}} \xrightarrow[(M+N)\to\infty]{} \mathcal{N}(0,1),$$

Beweis. Die folgenden Ausführungen orientieren sich an Kustosz, Müller und Wendler (2016). Zunächst werden aus den Zufallsvariablen $Z = (Z_1, ..., Z_{M+N})$ definiert

$$V_n := \prod_{k=1}^{K+1} \mathbb{1}\{\text{res}(\mu, Z_{n+k-1})(-1)^k > 0\} + \prod_{k=1}^{K+1} \mathbb{1}\{\text{res}(\mu, Z_{n+k-1})(-1)^{k+1} > 0\}$$

für alle $n = 1, ..., M+N-K$.

Die Zufallsvariablen $V_1, ..., V_{M+N-K}$ sind K-abhängig und identisch verteilt. Ferner sind sie als Bernoulli-Zufallsvariablen beschränkt und daher existieren auch ihre dritten Momente gleichmäßig. Fasst man für beliebige $M, N \in \mathbb{N}$ diese Zufallsvariablen als Folge auf, werden alle Voraussetzungen aus Satz 2.14 erfüllt und der zentrale Grenzwertsatz für K-abhängige Zufallsvariablen kann angewendet werden.

Zunächst berechnet man den Erwartungswert für beliebiges n:

$$\begin{aligned} E(V_n) &= E\left(\prod_{k=1}^{K+1} \mathbb{1}\{\mathrm{res}(\mu, Z_{n+k-1})(-1)^k > 0\} + \prod_{k=1}^{K+1} \mathbb{1}\{\mathrm{res}(\mu, Z_{n+k-1})(-1)^{k+1} > 0\}\right) \\ &= \prod_{k=1}^{K+1} P(\mathrm{res}(\mu, Z_{n+k-1})(-1)^k > 0) + \prod_{k=1}^{K+1} P(\mathrm{res}(\mu, Z_{n+k-1})(-1)^{k+1} > 0) \\ &= \left(\frac{1}{2}\right)^{K+1} + \left(\frac{1}{2}\right)^{K+1} = \left(\frac{1}{2}\right)^{K}. \end{aligned}$$

In der letzten Zeile geht vor allem die Annahme an den Median ein. Ferner sei nun A aus Satz 2.15 zu bestimmen. Dazu betrachte man vorerst folgende Rechnung:

$$\begin{aligned} \text{Für beliebiege } i \in \mathbb{N}_0 : \ & \frac{1}{p}\sum_{h=1}^{p} A_{i+h} \\ &= \frac{1}{p}\sum_{h=1}^{p}\left(\mathrm{Var}(V_{i+h+K}) + 2\sum_{j=1}^{K}\mathrm{Cov}(V_{i+h+K-j}, V_{i+h+K})\right) \\ &= \mathrm{Var}(V_{i+1+K}) + 2\sum_{j=1}^{K}\mathrm{Cov}(V_{i+1+K-j}, V_{i+1+K}) \end{aligned}$$

wobei die identische Verteilung der $V_1, ..., V_{M+N-K}$ in der letzten Gleichheit eingeht. Insbesondere ist der Ausdruck für jedes $i \in \mathbb{N}_0$ gleich und unabhängig von p. Also hat A folgende Gestalt:

$$A = \mathrm{Var}(V_1) + 2\sum_{j=1}^{K}\mathrm{Cov}(V_{K+1-j}, V_{K+1}).$$

Beide Summanden werden nun nacheinander bestimmt.

$$\begin{aligned} \mathrm{Var}(V_1) &= E\left(\left((V_1 - \left(\frac{1}{2}\right)^K\right)^2\right) \\ &= E(V_1^2) - 2E(V_1)\left(\frac{1}{2}\right)^K + \left(\frac{1}{2}\right)^{2K} \\ &= \left(\frac{1}{2}\right)^K - \left(\frac{1}{2}\right)^{2K} = \left(\frac{1}{2}\right)^K\left(1 - \left(\frac{1}{2}\right)^K\right). \end{aligned}$$

Ferner gilt für $j = 1, ..., K$

$$\begin{aligned}\mathrm{Cov}(V_{K+1-j}, V_{K+1}) &= E\left(\left(V_{K+1-j} - \left(\frac{1}{2}\right)^K\right)\left(V_{K+1} - \left(\frac{1}{2}\right)^K\right)\right) \\ &= E(V_{K+1-j}V_{K+1}) - E(V_{K+1-j})\left(\frac{1}{2}\right)^K - E(V_{K+1})\left(\frac{1}{2}\right)^K + \left(\frac{1}{2}\right)^{2K} \\ &= \left(\frac{1}{2}\right)^{K+j} - \left(\frac{1}{2}\right)^{2K}.\end{aligned}$$

Insgesamt folgt daraus für A

$$\begin{aligned}A &= \left(\frac{1}{2}\right)^K\left(1 - \left(\frac{1}{2}\right)^K\right) + 2\sum_{j=1}^{K}\left(\left(\frac{1}{2}\right)^{K+j} - \left(\frac{1}{2}\right)^{2K}\right) \\ &= \left(\frac{1}{2}\right)^K\left(1 - \left(\frac{1}{2}\right)^K\right) + \left(\frac{1}{2}\right)^K\sum_{j=0}^{K-1}\left(\frac{1}{2}\right)^j - K\left(\frac{1}{2}\right)^{2K-1}.\end{aligned}$$

Die Verwendung der geometrischen Summenformel ergibt weiterhin:

$$\begin{aligned}&= \left(\frac{1}{2}\right)^K\left(1 - \left(\frac{1}{2}\right)^K + \frac{1 - \left(\frac{1}{2}\right)^K}{1 - \frac{1}{2}} - K\left(\frac{1}{2}\right)^{K-1}\right) \\ &= \left(\frac{1}{2}\right)^K\left(1 - \left(\frac{1}{2}\right)^K + 2 - 2\left(\frac{1}{2}\right)^K - K\left(\frac{1}{2}\right)^{K-1}\right) \\ &= \left(\frac{1}{2}\right)^K\left(3 - 3\left(\frac{1}{2}\right)^K - K\left(\frac{1}{2}\right)^{K-1}\right).\end{aligned}$$

Satz 2.14 impliziert schließlich die Behauptung

$$\frac{1}{\sqrt{M+N}}\frac{\sum_{i=1}^{M+N-K}(V_i - E(V_i))}{\sqrt{A}} \xrightarrow[(M+N)\to\infty]{} \mathcal{N}(0,1).$$

□

Analog zur vollen Dreier-Tiefe lässt sich mithilfe der asymptotischen Verteilung ein asymptotisches Testverfahren für die vereinfachte Tiefe konstruieren.

Satz 2.16 (Zweistichproben-Relevanz-Test mit vereinfachter Tiefe).
Seien $X_1, ..., X_M$ und $Y_1, ..., Y_N$ beliebige, stetige Zufallsvariablen, die den Generalannahmen aus Abschnitt 2 genügen. Man beachte die Notation $Z = (X, Y)$ für die zu-

sammengefassten Stichproben. Sei das Hypothesenpaar $H_0 : |\mu_1 - \mu_2| \leq \delta$ *gegen* $H_1 : |\mu_1 - \mu_2| > \delta$ *mit Annahmebereich* $\Theta_0 := \{\mu \in \mathbb{R}^2 \mid |\mu_1 - \mu_2| \leq \delta\}$ *gegeben. Folgende Entscheidungsregel liefert dann einen asymptotischen Test zum Niveau* α*:*

$$\textit{Man verwerfe } H_0, \textit{ falls } \sup_{\mu \in \Theta_0} T^S_{M+N;(K)}(\mu, Z) < u_\alpha,$$

wobei u_α *das* α*-Quantil der Standardnormalverteilung ist.*

Beweis. Komplett analog zu Satz 2.12. □

Im Gegensatz zur vollen $(K+1)$-Tiefe, sind für beliebige $K \in \mathbb{N}$ die asymptotischen Verteilungen der vereinfachten $(K+1)$-Tiefe bekannt.

Geschwindigkeit der Asymptotik der vereinfachten Dreier-Tiefe

Wie bei der vollen Dreier-Tiefe können auch bei der vereinfachten Dreier-Tiefe notwendige Bedingungen für den Stichprobenumfang in Abhängigkeit vom Signifikanzniveau α für die Asymptotik formuliert werden. Dabei betrachtet man analog die normierte Teststatistik für die vereinfachte Dreier-Tiefe

$$\sqrt{M+N-2}\frac{d^S_{M+N}(\mu, z) - \frac{1}{4}}{\sqrt{\frac{5}{16}}}$$

und minimiert diesen Ausdruck, indem $d^S_{M+N}(\mu, z) = 0$ gesetzt wird. Anschließend untersucht man, welche Stichprobenumfänge $M+N$ überhaupt eine Unterschreitung des α-Quantils u_α einer Standnormalverteilung ermöglichen können:

$$-\sqrt{M+N-2}\frac{\frac{1}{4}}{\sqrt{\frac{5}{16}}} \overset{!}{\leq} u_\alpha \Rightarrow M+N \overset{!}{\geq} 5u_\alpha^2 + 2.$$

In der nachfolgenden Tabelle 4 sind für typische Signifikanzniveaus die zur nächsten natürlichen Zahl aufgerundeten, notwendigen Stichprobenumfänge angegeben, sodass eine Asymptotik gewährleistet werden kann:
Verglichen mit Tabelle 3 sind die notwendigen Stichprobenumfänge deutlich größer.

Signifikanzniveau α	0.05	0.01	0.001
Stichprobenumfang für $M+N$	16	30	50

Tabelle 4: notwendiger Stichprobenumfang $M+N$ für ein gegebenes Signifikanzniveau α bei der vereinfachten Dreier-Tiefe

2.3.3 Numerische Berechnung des Supremums beim Testverfahren

Die Berechnung des Supremums über der Menge $\{\mu \in \mathbb{R}^2 : |\mu_1 - \mu_2| \leq \delta\}$ birgt einige Hürden mit sich. Eine simple Diskretisierung dieser Menge durch ein Gitter ist wegen ihrer Unbeschränktheit noch nicht ausreichend, da trotzdem noch unendlich viele Parameter überprüft werden müssten. Man muss sich auf beschränkte Regionen dieser Menge zurück ziehen, in denen man mit hoher Sicherheit vermuten kann, dass dort das Supremum angenommen wird. Dies wird mit Bereichsschätzungen für μ_1 und μ_2 aus den Daten durchgeführt. Sei $[c_l^i, c_u^i]$ eine Konfidenzbereichschätzung zum Niveau $1 - \frac{\alpha}{L}$ für μ_i mit $i = 1, 2$ und $L > 0$. Für hinreichend große $K \geq 0$ sei

$$C^i := \{c_l^i, c_l^i + \varepsilon, ..., c_u^i - \varepsilon, c_u^i\} \text{ mit } \varepsilon = \frac{c_u^i - c_l^i}{K}$$

das Konfidenzgitter für μ_i. L skaliert dabei die Länge des Konfidenzgitters und K die Feinheit. Nun wird folgendes Konfidenzgitter für $\mu = (\mu_1, \mu_2)$ gebildet

$$C = \{\mu \in C^1 \times C^2 \mid |\mu_1 - \mu_2| \leq \delta\},$$

wobei δ der entsprechende Relevanzparameter ist. Dadurch gewinnt man eine Näherung für $\Theta_0 = \{\mu \in \mathbb{R}^2 \mid |\mu_1 - \mu_2| \leq \delta\}$ und erwartet

$$\sup_{\mu \in C} T_{M+N}(\mu, z) \approx \sup_{\mu \in \Theta_0} T_{M+N}(\mu, z)$$

(bzw. $T^S_{M+N}(\mu, z)$ statt $T_{M+N}(\mu, z)$). Findet man einen Gitterpunkt, in dem die normierte Tiefe bereits den kritischen Wert q_α bzw. u_α überschreitet, so kann man die Suche nach dem Supremum direkt abbrechen, da das Supremum mindestens genauso groß sein wird und behält die Nullhypothese bei.
Liegt eine Situation vor, in der die Nullhypothese verworfen werden soll, muss je-

der Gitterpunkt untersucht werden. Unterschreiten alle Werte der Gitterpunkte den kritischen Wert, so geht man davon aus, dass das Supremum auch den kritischen Wert nicht überschreitet. In diesem Fall lehnt man die Nullhypothese ab. Dementsprechend ist der Rechenaufwand im Allgemeinen größer, wenn die Nullhypothese abgelehnt werden soll. Hier ist zu beachten, dass die Parameter K, L hinreichend groß gewählt werden müssen, da ansonsten das Supremum verpasst wird und man eventuell fälschlicherweise die Nullhypothese ablehnt. Die gewählte Bereichsschätzung kann zum Beispiel über ein Konfidenzintervall für den Erwartungswert μ_i für $i = 1, 2$ einer Normalverteilung bei unbekannter Varianz wie in Formel (S) erfolgen. In Formel (S) wird das Konfidenzintervall für die erste Stichprobe $X_1, ..., X_M$ angegeben:

$$[c_l^1, c_u^1] = \left[\overline{X.} - t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X, \overline{X.} + t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X\right]. \tag{S}$$

Eine analoge Gestalt besitzt das Konfidenzintervall $[c_l^2, c_u^2]$ für die zweite Stichprobe $Y_1, ..., Y_N$. Statt dem Stichprobenmittel zur Schätzung des Erwartungswerts kann auch der Median als Lageschätzer verwendet werden, da der Median eine konsistente Schätzung für den Lageparameter liefert und das Stichprobenmittel im Fall von Realisationen cauchyverteilter Zufallsvariablen vor allem nicht konsistent für den Lageparameter μ_i ist. Formel (M) stellt das beschriebene Konfidenzintervall für eine Stichprobe $X_1, ..., X_M$ dar:

$$[c_l^1, c_u^1] = \left[\text{med}(X) - t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X, \text{med}(X) + t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X\right]. \tag{M}$$

Um die Güteeigenschaften der Tests auf Tiefe basierend möglichst exakt zu bestimmen, fließt das Wissen über die wahre Verteilung in der Simulationsstudie ein. In der Praxis können nur Vermutungen über die wahre Verteilung aufgestellt werden, sodass die Frage bleibt, welche Konfidenzgitter im Allgemeinen bei der Suche des Supremums verwendet werden sollten.

2.3.4 Idee eines Vorzeichentests

In diesem Unterabschnitt soll die Frage aufgegriffen werden, warum nicht die Idee eines üblichen Vorzeichentests für zwei Stichproben verwendet wird. Analog zur Datentiefe betrachtet man die Vorzeichen der Residuen $\text{res}(\mu, z)$. Man betrachte:

$$\frac{1}{M+N} \sum_{n=1}^{M+N} \left(\mathbb{1}\{\text{res}(\mu, z_n) > 0\} - \frac{1}{2} \right).$$

Dieser Ausdruck bestimmt den relativen Anteil der positiven Vorzeichen von allen Daten. Anschließend wird durch $\frac{1}{2}$ zentriert. Ist diese Differenz für einen gewählten Parameter $\mu = (\mu_1, \mu_2)$ im Nicht-Relevanzbereich nahe bei 0, so spricht das für ähnliche Anteile von positiven und negativen Vorzeichen. Das wiederum bedeutet, dass die gewählten Mittelwerte (μ_1, μ_2) zu den Daten passen, weswegen man die Nullhypothese nicht verwerfen sollte. Für den wahren Parameter $\mu = (\mu_1, \mu_2)$ gilt nach Standardisierung der Varianz mit dem Zentralen Grenzwertsatz die asymptotische Normalität der folgenden Teststatistik

$$T^B_{M+N}(\mu, Z) = \sqrt{M+N} \left(\frac{\frac{1}{M+N} \sum_{n=1}^{M+N} \mathbb{1}\{\text{res}(\mu, Z) > 0\} - \frac{1}{2}}{\frac{1}{2}} \right) \xrightarrow[(M+N)\to\infty]{} \mathcal{N}(0,1).$$

Das motiviert folgende Entscheidungsregel für das Hypothesenpaar (H):

> Man verwerfe H_0,
>
> falls $\sup\limits_{|\mu_1-\mu_2|\leq\delta} T^B_{M+N}(\mu, z) < u_{\alpha/2}$ oder $\inf\limits_{|\mu_1-\mu_2|\leq\delta} T^B_{M+N}(\mu, z) > u_{1-\alpha/2}$,

wobei u_α das α-Quantil der Standnormalverteilung sei.
Das gravierende Problem bei diesem Ansatz ist, dass das Supremum und Infimum bereits unabhängig von dem echten Parameter $\mu = (\mu_1, \mu_2)$ bestimmt sind:

$$\sup_{|\mu_1-\mu_2|\leq\delta} T^B_{M+N}(\mu, z) = \sqrt{M+N} \text{ und } \inf_{|\mu_1-\mu_2|\leq\delta} T^B_{M+N}(\mu, z) = -\sqrt{M+N}.$$

Um das einzusehen, wählt man für das Supremum bzw. Infimum zum Beispiel den Parameter $\tilde{\mu} = (\max(z), \max(z))$ bzw. $\tilde{\mu} = (\min(z), \min(z))$ und erhält die angege-

ben Werte, welche gerade dem Maximum bzw. Minimum von $T^B_{M+N}(\mu, z)$ entsprechen. Somit würde der Test die Nullhypothese unabhängig von den Daten und den wahren Parametern nie verwerfen können.

2.4 Übersicht über Verteilungsklassen der Fehlerterme

Neben normalverteilten Fehlern können noch andere Verteilungen betrachtet und die Güteeigenschaften der Testverfahren in diesen Fällen studiert werden.

2.4.1 Cauchyverteilung

Eine der Normalverteilung sehr ähnelnde, in ihren Eigenschaft allerdings dennoch sich unterscheidende Verteilung ist die Cauchyverteilung. Sie lässt sich wie folgt definieren (Büning und Trenkler 1994, S.26).

Definition 2.17 (Cauchyverteilung).
Eine Zufallsvariable Z heißt cauchyverteilt mit Lokationsparameter μ und Streuungsparameter $\gamma > 0$, kurz: $Z \sim Cau(\mu, \gamma)$, falls die Verteilung P_Z von Z absolut-stetig zum Lebesgue-Maß mit folgender Lebesgue-Dichte ist:

$$f(x) = \frac{\gamma}{\pi(\gamma^2 + (x-\mu)^2)} \text{ für } x \in \mathbb{R}.$$

Es ist einfach einzusehen, dass f der Dichte einer Zufallsvariable entspricht. Da $\gamma > 0$ gilt und der Nenner aus einer Summe quadratierter Terme besteht, ist $f(x) \geq 0$ für alle $x \in \mathbb{R}$. Die Normalitätseigenschaft lässt mithilfe der elementare Substitutionsregel für Integrale mit $y = \frac{x-\mu}{\gamma}$ nachrechnen:

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \frac{1}{\gamma\pi}\int_{-\infty}^{\infty} \frac{1}{\left(1 + (\frac{x-\mu}{\gamma})^2\right)}\,dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{(1+y^2)}\,dy \\ &= \frac{1}{\pi}\Big[\arctan(y)\Big]_{-\infty}^{\infty} = \frac{1}{\pi}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1.\end{aligned}$$

Man bezeichnet μ als den Lageparameter der Cauchyverteilung, da er der Median ist. Setzt man $\mu = 0$, so erfüllen die Cauchyverteilungen die geforderten Eigenschaften an die Fehler $E_1, ..., E_{M+N}$ aus der Generalannahme. Weiterhin nennt man γ

den Skalenparameter der Cauchyverteilung, da er die Breite der Dichte skaliert.
Man kann die Cauchyverteilung als Gegenstück zur Normalverteilung auffassen. Beide Verteilungen ähneln sich zwar in ihrer symmetrischen Struktur und besitzen einen Träger über $\mathbb{R}$. Allerdings besitzt die Cauchyverteilung sogenannte *heavy tails*, produziert also extremere Werte als die Normalverteilung und liefert dadurch bei Robustheitsanalysen unterschiedliche Resultate.
Für spätere Vergleiche der Tests ist der Fakt wichtig, dass die Momente der Cauchyverteilung für beliebige Parameter μ, γ nicht existieren (siehe Büning und Trenkler (1994), S.27), d.h.

$$E(|X|^k) = \infty \text{ für alle } k \in \mathbb{N}_{\geq 1}, \text{ falls } X \sim Cau(\mu, \gamma).$$

Dadurch existiert kein Varianzbegriff, den aber gerade der t-Test aus Abschnitt 2.2.3 benötigt. Die damit verbundenen Probleme werden in Abschnitt 3.3.3 in der Simulationsstudie dargestellt.

2.4.2 Kontaminationen

Eine weitere Möglichkeiten Stichproben mit extremen Werten für Robustheitsuntersuchungen zu produzieren ist die Verteilung mit einem Kontaminationsparameter $\varepsilon \in (0, 1)$ zu verunreinigen, wie es bei der Bestimmung und Verwendung der kontaminierten Normalverteilung gemacht werden kann (siehe Büning und Trenkler (1994), S.24 und S.294ff). Der Parameter ε gibt dabei die Intensität an, wie stark die Verteilung von Ausreißern befallen ist. Die hier verwendete Idee ist die Bildung einer Konvexkombination aus einem Grund-Wahrscheinlichkeitsmaß P und einem Wahrscheinlichkeitsmaß Q, das mit Intensität ε das Maß P kontaminiert:

Satz 2.18 (Kontaminierte Verteilungen).
Seien P, Q zwei Wahrscheinlichkeitsmaße auf einem Messraum $(\Omega, \mathcal{A})$. Dann bildet die Q-ε-Kontamination mit P und $\varepsilon \in (0, 1)$

$$P_{Q,\varepsilon} := (1 - \varepsilon)P + \varepsilon Q$$

ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathcal{A})$.

Beweis. $P_{Q,\varepsilon}$ ist eine wohldefinierte Abbildung auf dem Definitionsbereich $\mathcal{A}$, da sie durch eine Linearkombination von zwei Abbildungen der Form $\mathcal{A} \to [0,1]$ definiert ist. Weiterhin können die Anforderungen an ein Wahrscheinlichkeitsmaß elementar nachgerechnet werden. Für $\Omega \in \mathcal{A}$ gilt

$$P_{Q,\varepsilon}(\Omega) = (1-\varepsilon)P(\Omega) + \varepsilon Q(\Omega) = (1-\varepsilon) + \varepsilon = 1.$$

Für eine Folge disjunkter Mengen $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}$ folgt

$$\begin{aligned} P_{Q,\varepsilon}\left(\bigcup_{n\in\mathbb{N}} A_n\right) &= (1-\varepsilon)P\left(\bigcup_{n\in\mathbb{N}} A_n\right) + \varepsilon Q\left(\bigcup_{n\in\mathbb{N}} A_n\right) \\ &= (1-\varepsilon)\sum_{n\in\mathbb{N}} P(A_n) + \varepsilon \sum_{n\in\mathbb{N}} Q(A_n) \\ &= \sum_{n\in\mathbb{N}} [(1-\varepsilon)P(A_n) + \varepsilon Q(A_n)] = \sum_{n\in\mathbb{N}} P_{Q,\varepsilon}(A_n). \end{aligned}$$

In der ersten Rechnung verwendet man, dass P und Q normiert sind, während man in der zweiten Rechnungen die σ-Additivität der Maße P und Q benötigt. □

Die Wohldefiniertheit von $P_{Q,\varepsilon}$ wird dadurch gewährleistet, dass P und Q Wahrscheinlichkeitsmaße auf dem gleichen Grundraum Ω mit einer passenden σ-Algebra $\mathcal{A}$ sind. Da in der Praxis meist auf $\Omega = \mathbb{R}^n$ gearbeitet wird, findet man aber in der Regel einen gemeinsamen Messraum.
In der Simulationsstudie wird eine Realisation einer Zufallsvariablen mit Verteilung $P_{Q,\varepsilon}$ durch ein vorgeschaltetes Bernoulli-Experiment zum Parameter ε bestimmt. Diese entscheidet mit Wahrscheinlichkeit $(1-\varepsilon)$ bzw. ε, ob eine Realisation einer Zufallsvariablen mit Verteilung P bzw. Q simuliert wird.

3 Simulationsstudie und Auswertung

Die Qualität eines statistischen Testverfahrens wird an seiner Gütefunktionen $G_\varphi(\vartheta)$ gemessen. Für ein Testverfahren mit Entscheidungsregel $\varphi(X)$ und Parameter $\vartheta \in \Theta$ aus dem Parameterraum Θ ist sie definiert durch den Erwartungswert der Entscheidungsregel zum Parameter ϑ (vgl. Georgii 2002, S.247f)

$$G_\varphi(\vartheta) := E_\vartheta(\varphi).$$

Entscheidungsregeln können im nicht-randomisierten Fall nur Werte aus $\{0, 1\}$ annehmen. Somit besitzt die Gütefunktion in diesem Fall folgende Gestalt

$$G_\varphi(\vartheta) = P_\vartheta(\varphi = 1).$$

Die Gütefunktion beschreibt also die Wahrscheinlichkeit, dass ein Testverfahren die Nullhypothese ablehnt, unter der Annahme, dass ϑ der wahre Parameter ist. Aus ihr lässt sich für einen festen Stichprobenumfang der Fehler erster Art und zweiter Art ablesen. Für Parameter im Annahmebereich $\vartheta \in \Theta_0$ beschreibt die Gütefunktion den Fehler erster Art. Für Parameter im Ablehnungsbereich $\vartheta \in \Theta_1$ beschreibt sie die Gegenwahrscheinlichkeit des Fehlers zweiter Art.
Ein Testverfahren soll in erster Linie die Bedingung erfüllen, dass der Fehler erster Art unter dem Signifikanzniveau α liegt. Das heißt, dass die Gütefunktion im Bereich Θ_0 nicht das Signifikanzniveau α überschreiten darf. Eventuell kann diese Bedingung erst bei höherem Stichprobenumfang erfüllt werden. Weiterhin ist es wünschenswert, dass ein Testverfahren für Parameter im Ablehnungsbereich Θ_1 möglichst hohe Werte für ihre Gütefunktion besitzt, also vor allem möglichst nah an 1 ist.
Die Gütefunktion lässt sich im Allgemeinen nicht explizit bestimmen. In so einem Fall wird sie durch eine Simulation approximativ durch das *Gesetz der großen Zahlen* (GdgZ) berechnet (Georgii 2002, S.114ff). Man betrachtet eine hinreichend hohe Anzahl S von Testentscheidungen als Realisation der Entscheidungsregeln. Nach dem GdgZ entspricht die Realisation des Stichprobenmittels der Testentscheidun-

gen $\varphi\left(x_1^{(s)}, ..., x_{M+N}^{(s)}\right)$ über alle $s = 1, ..., S$ ungefähr dem wahren Erwartungswert:

$$G_\varphi(\vartheta) = E_\vartheta(\varphi) \overset{\text{GdgZ}}{\approx} \frac{1}{S} \sum_{s=1}^{S} \varphi\left(x_1^{(s)}, ..., x_{M+N}^{(s)}\right).$$

Dabei beschreibt $x^{(s)} = \left(x_1^{(s)}, ..., x_{M+N}^{(s)}\right)$ die s-te Stichprobe des s-ten durchgeführten Tests der Simulation. Die Voraussetzungen des GdgZ sind erfüllt, wenn die Testentscheidungen durch unabhängige, identische Zufallszahlen erzeugt werden. Außerdem müssen Bedingungen für die Integrierbarkeit, also die Existenz des Erwartungswerts, gesichert sein. Die Integrierbarkeitsbedingung $E_\vartheta|\varphi| < \infty$ ist erfüllt, da jede Entscheidungsregel φ als $\{0, 1\}$-wertige Zufallsvariable beschränkt ist. Die rechte Seite der obigen Approximation kann durch wiederholte Simulation von Daten und Anwendungen der Entscheidungsregel bestimmt werden. Die Wahl von S variiert in der Simulationsstudie je nach Rechenaufwand der Testverfahren und wird im nachfolgenden Text an den jeweiligen Stellen explizit angegeben.

3.1 Vorgehensweise und Ziele

In der folgenden Simulationsstudie werden zunächst die Zweistichproben-Relevanz-t-Tests miteinander verglichen. Als erstes wird der Idealfall betrachtet, in dem alle Voraussetzungen erfüllt sind. Dabei wird sich ergeben, dass die t-Tests für hohe Stichprobenumfänge sehr ähnliche Ergebnisse liefern. Daher werden die t-Tests auch unter sehr kleinen Stichprobenumfängen verglichen.

Ferner wird das Verhalten der t-Tests unter nicht erfüllten Voraussetzungen untersucht. Ein Vergleichspunkt werden falsche Annahmen an die Varianz sein. Das interessiert vor allem, da einer der t-Test ein von der Varianz abhängiges Hypothesenpaar besitzt. Weiterhin wird untersucht, wie sich die t-Tests bei cauchyverteilten Fehlern verhalten, da diese Verteilungsklasse anders als die vorauszusetzende Normalverteilung extremere Werte produziert. Dadurch lassen sich die Robustheit der t-Tests und ihr Verhalten bei falschen Modellannahmen analysieren.

Im zweiten Teil der Simulationsstudie werden die Testverfahren mit Verwendung der Datentiefe untersucht. Anfangs wird auf die numerischen Schwierigkeiten bei

der Suche des Supremums eingegangen. Anschließend werden die Tests mit voller und vereinfachter Dreier-Tiefe mit den t-Tests bei Normal- und Cauchyverteilung verglichen. Der Vergleich wird ebenso für unterschiedliche Stichprobenumfänge durchgeführt, um die Asymptotik der Tests zu untersuchen.
Abschließend werden einige Vergleiche der Test bei kontaminierten Datensätzen betrachtet.

3.2 Vereinfachung der grafischen Darstellung

Da bei Zweistichproben-Tests zwei Parameter vorliegen, besitzen ihre Gütefunktionen einen zweidimensionalen Definitionsbereich. Exemplarisch wird in der Abbildung 3 die Gütefunktion des t-Tests aus Abschnitt 2.2.3 für einen Stichprobenum-

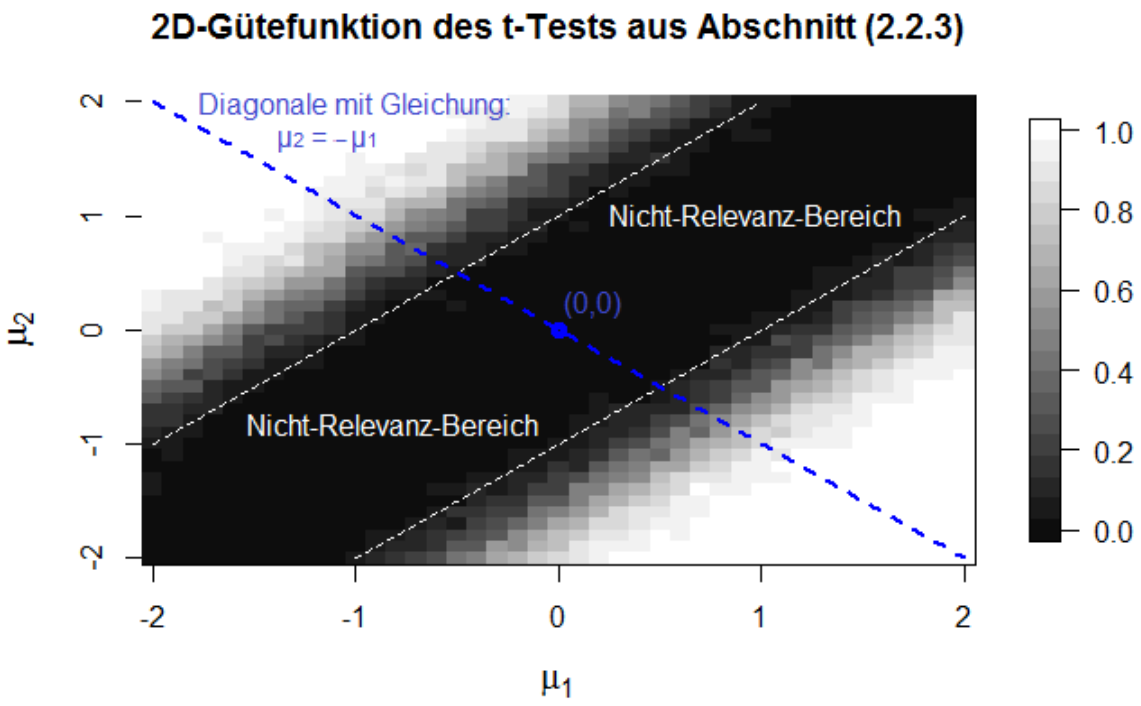


Abbildung 3: Darstellung der Gütefunktion des t-Tests aus Abschnitt 2.2.3 auf dem Parameterbereich $[-2, 2]^2$

fang von $M = N = 20$ mit Relevanzparameter $\delta = 1$ zum Signifikanzniveau $\alpha = 0.05$ für normalverteilte Fehler E_i für $i = 1, ..., M + N$ mit $\sigma^2 = 1$ dargestellt. Die Anzahl an Wiederholungen pro Gitterpunkt beträgt $S = 100$, wobei das Gitter in 0.01er

Schritten durchlaufen wird. Der Nicht-Relevanzbereich Θ_0 ist der durch die gestrichelten, schwarzen Linien begrenzte Fläche.
Die Nachteile dieser Darstellung sind einerseits die zeitintensive Berechnungen und andererseits sind Vergleiche mehrerer Gütefunktionen weniger übersichtlich. Daher wird folgender Ansatz für die Darstellung verwendet: Die Abbildung 3 zeigt, dass die Gütefunktion symmetrische Strukturen aufweist. Statt die vollständige Gütefunktion zu betrachten, wird sie auf der Gerade mit Gleichung $\mu_2 = -\mu_1$ untersucht. Eindimensionale Projektionen der Gütefunktionen auf dieser Geradengleichung ermöglichen es, mehrere Gütefunktionen ohne Informationsverlust in einer Grafik zu vergleichen und sparen viel Rechenzeit ein. Dadurch können auch die Simulationen auf der Geraden feiner durchgeführt werden.

3.3 Vergleich der t-Tests

Die nachfolgenden Untersuchungen werden stets auf der Geraden $\mu_2 = -\mu_1$ mit einer Feinheit in 0.01er Schritten mit 4000 Simulationen bei Normalverteilungen und 1000 Simulationen bei Cauchyverteilungen (da sich im Gegensatz zur Normalverteilung ab ca. 1000 keine deutlichen Verbesserungen mehr ergeben) durchgeführt.

3.3.1 Vergleich der t-Tests unter Normalverteilung

In der Abbildung 4 werden für verschiedene Stichprobenumfänge M, N die Gütefunktionen der drei t-Tests aus Abschnitt 2.2 mit Relevanzparameter $\delta = 1$ für normalverteilte Fehler E_i für $i = 1, ..., M+N$ mit $\sigma^2 = 1$ dargestellt. Dazu wird das Signifikanzniveau $\alpha = 0.05$ durch eine horizontale und der Annahmebereich (Projektion auf $[-\frac{1}{2}, \frac{1}{2}]$ auf die Diagonale) durch vertikal gestrichelte Linien dargestellt. Aus der Grafik lässt sich entnehmen, dass alle Testverfahren, wie theoretisch bereits gezeigt, das Signifikanzniveau α einhalten. Jedoch schöpft lediglich der t-Test mit varianzabhängiger Hypothese aus Abschnitt 2.2.3 das Signifikanzniveau vollständig aus, das heißt, er besitzt am Rand des Annahmebereichs exakt die Güte α. Die anderen beiden Testverfahren sind *konservativ* (nicht unverfälscht) - sie besitzen also für kleinere Lageunterschiede einen Fehler zweiter Art größer als 0.95. Der t-Test auf der Konfidenzintervall-Exklusion basierend aus Abschnitt 2.2.1 und der t-Test

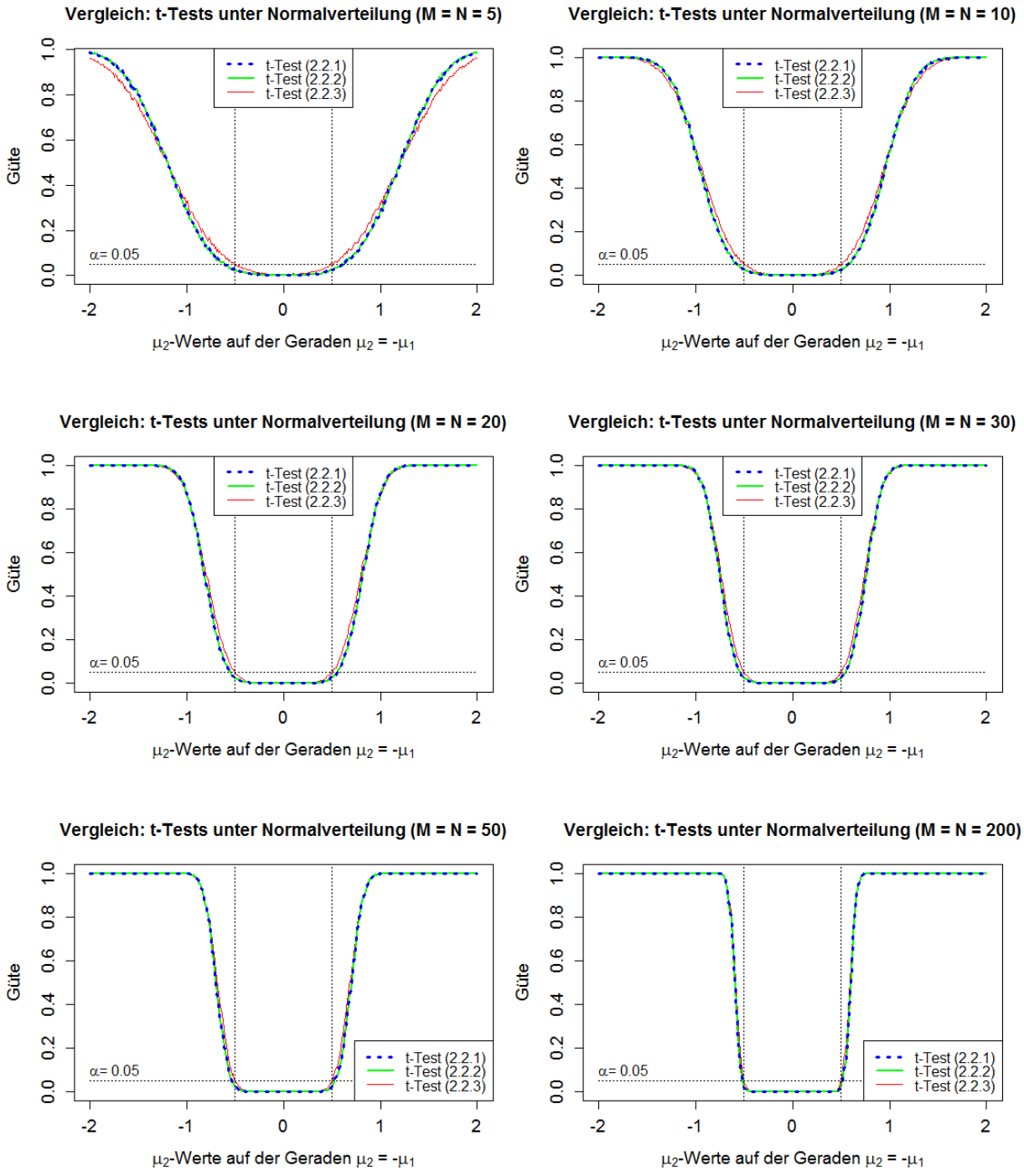


Abbildung 4: Gütefunktion der t-Tests für verschiedene Stichprobenumfänge

auf multiplen Hypothesen basierend aus Abschnitt 2.2.2 besitzen für alle Stichprobenumfänge, wie im Theorieteil gezeigt, die gleiche Güte. Die beiden äquivalenten t-Tests besitzen im Ablehnungsbereich für geringere Lageunterschiede eine niedrigere Güte als der t-Test aus 2.2.3. Für hohe Lageunterschiede können sie den Güteverlust aber aufholen und besitzen dann eine höhere Güte als der t-Test aus Abschnitt 2.2.3. Keiner der drei t-Test besitzt demnach gleichmäßig eine höhere Güte als die anderen, wodurch unter ihnen *keine Optimalitätsaussage* formuliert werden kann.
Darüber hinaus erkennt man, dass sich die Testverfahren mit erhöhtem Stichprobenumfang verbessern und die Unterschiede unter ihnen immer geringer werden. Das deutet auf die *Konsistenz* der Testverfahren hin.
Kann man die Varianz relativ genau einschätzen, liegt ein sehr geringer Stichprobenumfang vor und deutet der Sachkontext der Daten nicht auf besonders hohe Lageunterschiede hin, eignet sich der dritte t-Test gut. Erwartet man eher höhere Abweichungen, so sollte einer der ersten beiden t-Tests verwendet werden. Hier können bereits vorliegende Erfahrungen über die Abweichungen der Daten bei der Wahl des Testverfahrens eine Rolle spielen.
Sind die Stichproben relativ groß, unterscheiden sich die Tests kaum. In diesen Fällen ist die Verwendung einer der ersten beiden Tests am sichersten. Man beachte nämlich, dass die Simulationen unter der idealen Vorstellung durchgeführt wurden, dass die Varianz bekannt ist. Das ist in der Praxis in der Regel aber nicht gegeben. Die Sensibilität des dritten t-Tests bei falschen Varianzannahmen wird daher im Nachfolgenden untersucht und soll die Problematik bei Unsicherheit illustrieren.

3.3.2 Untersuchungen bei falscher Varianzannahme

Die Abbildung 5 stellt Gütefunktionen des t-Tests aus Abschnitt 2.2.3 jeweils mit richtiger, überschätzter und unterschätzter Varianz dar. Dabei seien wieder E_i für $i = 1, ..., M + N$ normalverteilt mit wahrer Varianz $\sigma^2 = 1$ und der Relevanzparameter bei $\delta = 1$. Durch falsche Varianzannahmen wird der Relevanzparameter δ falsch gewählt. Betrachtet man das Hypothesenpaar

$$H_0 : |\mu_1 - \mu_2| \leq \sigma\delta \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > \sigma\delta,$$

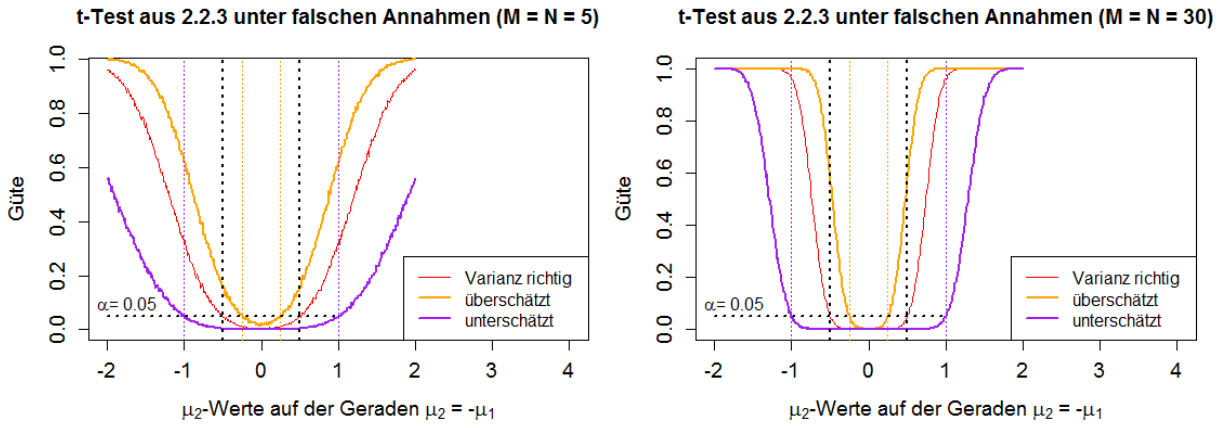


Abbildung 5: Gütefunktionen des t-Tests aus 2.2.3 bei falschen Varianzannahmen

so sieht man, dass δ in Abhängigkeit von der angenommenen Standardabweichung σ gewählt werden muss, um einen gewünschten Nicht-Relevanzbereich zu erhalten. Da $\sigma = 1$ gilt und man $\delta\sigma = 1$ als Nicht-Relevanzbereich erhalten möchte, wird $\delta = 1$ gewählt. Das heißt, folgendes Hypothesenpaar wird betrachtet

$$H_0 : |\mu_1 - \mu_2| \leq 1 \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > 1.$$

Überschätzt man die Standardbweichung mit $\sigma = 2$ (wie in Abbildung 5), so wird $\delta = \frac{1}{2}$ zu klein gewählt. Umgekehrt wählt man mit $\delta = 2$ zu groß, falls mit $\sigma = \frac{1}{2}$ die Standardabweichung unterschätzt (wie in Abbildung 5) wird. Dies bewirkt, dass falsche Hypothesenpaare unwissentlich untersucht werden. Im Fall der Überschätzung wird folgendes Hypothesenpaar untersucht:

$$H_0 : |\mu_1 - \mu_2| \leq \frac{1}{2} \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > \frac{1}{2}. \tag{O}$$

Im Fall einer Unterschätzung wird in Wahrheit dieses Hypothesenpaar untersucht:

$$H_0 : |\mu_1 - \mu_2| \leq 2 \quad \text{gegen} \quad H_1 : |\mu_1 - \mu_2| > 2. \tag{U}$$

Dies hat Einfluss auf die Entscheidungsregeln der Testverfahren, wie man in Abbildung 5 sehen kann. Eine Unterschätzung der Varianz führt zu einem konservativen Test, da die Größe des Annahmebereichs doppelt so groß wird. Überschätzungen

der Varianz führen zu Tests, die das Niveau nicht einhalten. Dementsprechend sind Überschätzungen der Varianz deutlich problematischer, da sie häufiger zu falschen Testentscheidungen führen als sie dürften. Bei erhöhtem Stichprobenumfang wird in der Grafik deutlich, dass bei falschen Annahmen die anderen, oben erwähnten Hypothesenpaare untersucht werden. Durch die Konsistenz der Tests nähern sich die Gütefunktionen zu den idealen Tests für die Hypothesenpaaren (O) und (U) an. In der Abbildung 5 soll dies durch die gestrichelten Linien in der jeweils passenden Farbe der Gütefunktion für den Annahmebereich verdeutlicht werden.
Unsicherheiten bezüglich der Varianz sollten daher ein wichtiges Entscheidungskriterium bei der Wahl der Testverfahren sein. Gerade bei hohen Stichprobenumfängen sind die Güteunterschied sehr gering (vgl. Abbildung 4), wodurch gegebenenfalls ein hohes Risiko eingegangen wird.

3.3.3 t-Test mit Konfidenzintervall-Exklusion mit gewichteten Seiten

Im Theorieteil wird eine Variation des t-Tests mit Konfidenzintervall-Exklusion (bzw. mit multiplen Hypothesen, da dieser äquivalent ist) erwähnt. Statt beide Seiten eines zweiseitigen Tests mit $\frac{\alpha}{2}$ beim Signifikanzniveau α zu adjustieren, sind auch beliebige $\alpha_1, \alpha_2 \geq 0$ zulässig, die $\alpha_1 + \alpha_2 = \alpha$ erfüllen. In der Abbildung 6 werden

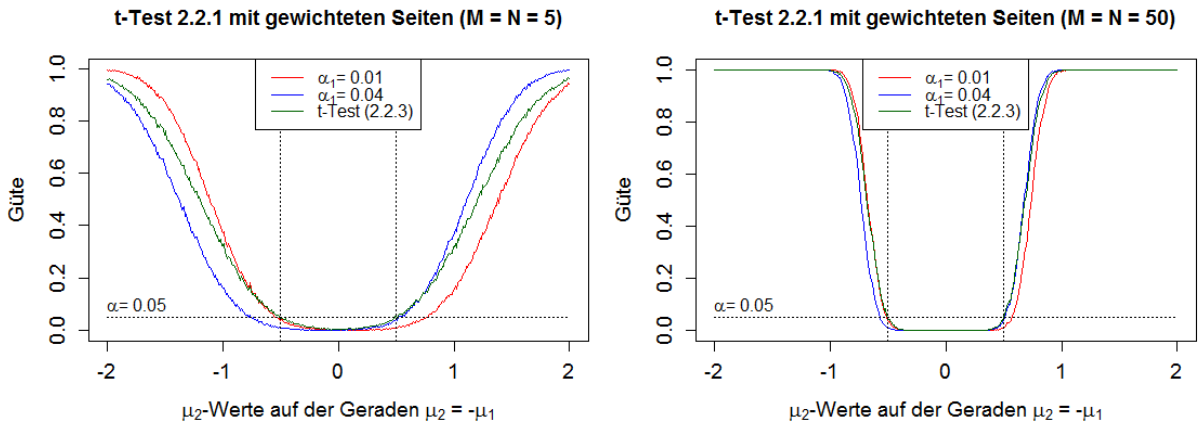


Abbildung 6: Gütefunktionen des t-Tests aus 2.2.1 mit gewichteten Seiten ($\alpha_1 = 0.01$ und $\alpha_2 = 0.04$ bzw. umgekehrt)

die Gütefunktionen für diese Testmethoden mit $\alpha_1 = 0.01$ (linke Intervallgrenze) und $\alpha_2 = 0.04$ (rechte Intervallgrenze) und umgekehrt für die Stichprobenumfänge

$M = N = 5,50$ dargestellt und mit der Gütefunktion des t-Tests aus Abschnitt 2.2.3 verglichen. Die geringer gewichtete Seite besitzt eine bessere Güte, während die höher gewichtete Seite eine niedrigere Güte besitzt. Je kleiner α_i für $i = 1, 2$ wird, desto eher nimmt die jeweilige auf dem Rand des Annahmebereichs das exakte Signifikanzniveau $\alpha = 0.05$ ein. Für $\alpha_i = 0$ gelangt man allerdings lediglich zu einem entarteten einseitigen Zweistichproben-t-Test aus Satz 2.4 und erhält damit kein neues Testverfahren. Ohnehin würde man vorher die Fragestellung bereits als Einstichproben-Problem auffassen, wenn man für ein $\alpha_i = 0$ setzt.

3.3.4 Vergleich der t-Tests unter Cauchyverteilung

In der Abbildung 7 werden die Gütefunktionen der t-Tests unter cauchyverteilten Fehlern E_i für $i = 1, ..., M+N$ für verschiedene Paare von Lageparametern μ analog zur Normalverteilung dargestellt. Der Skalenparameter beträgt dabei $\gamma = 1$ und der Relevanzparameter ist erneut $\delta = 1$.
Für den Relevanz-t-Test mit varianzabhängigen Hypothesenpaar aus Abschnitt 2.2.3 wird dabei eine Standardabweichung von $\sigma = 1$ gewählt. Problematisch ist hier der Varianzbegriff, da das zweite Moment der Cauchyverteilung nicht existiert und somit kein angemessenes Adäquat zur Varianz geschätzt werden kann. Eine Schätzung mittels der empirischen Varianz von cauchyverteilten Zufallsvariablen ist nicht konsistent. Daher soll im Vorfeld betont werden, dass die Annahme von $\sigma = 1$ nicht sinnvoll sein kann; es können aber generell keine mathematisch sinnvollen Annahmen getroffen werden.
Die Grafiken zeigen, dass alle t-Tests das Signifikanzniveau $\alpha = 0.05$ einhalten. Die t-Tests aus Abschnitt 2.2.1 und 2.2.2 liefern die gleichen und vor allem besseren Ergebnisse. Durch die unpassende Annahme an die Varianz für den dritten t-Test ergeben sich sehr niedrige Werte für die Gütefunktion.
Insgesamt sind alle drei Tests stark konservativ und erreichen selbst für hohe Unterschiede der Lageparameter keine Güte von 1. Es ist insbesondere auffällig, dass eine Erhöhung des Stichprobenumfangs keinen verbesserten Einfluss auf die Güte erzielt. Man könnte daher vermuten, dass die Tests unter Cauchyverteilung nicht konsistent sind. Insbesondere wird dies in Abbildung 8 verdeutlicht, in der aufgezeigt wird,

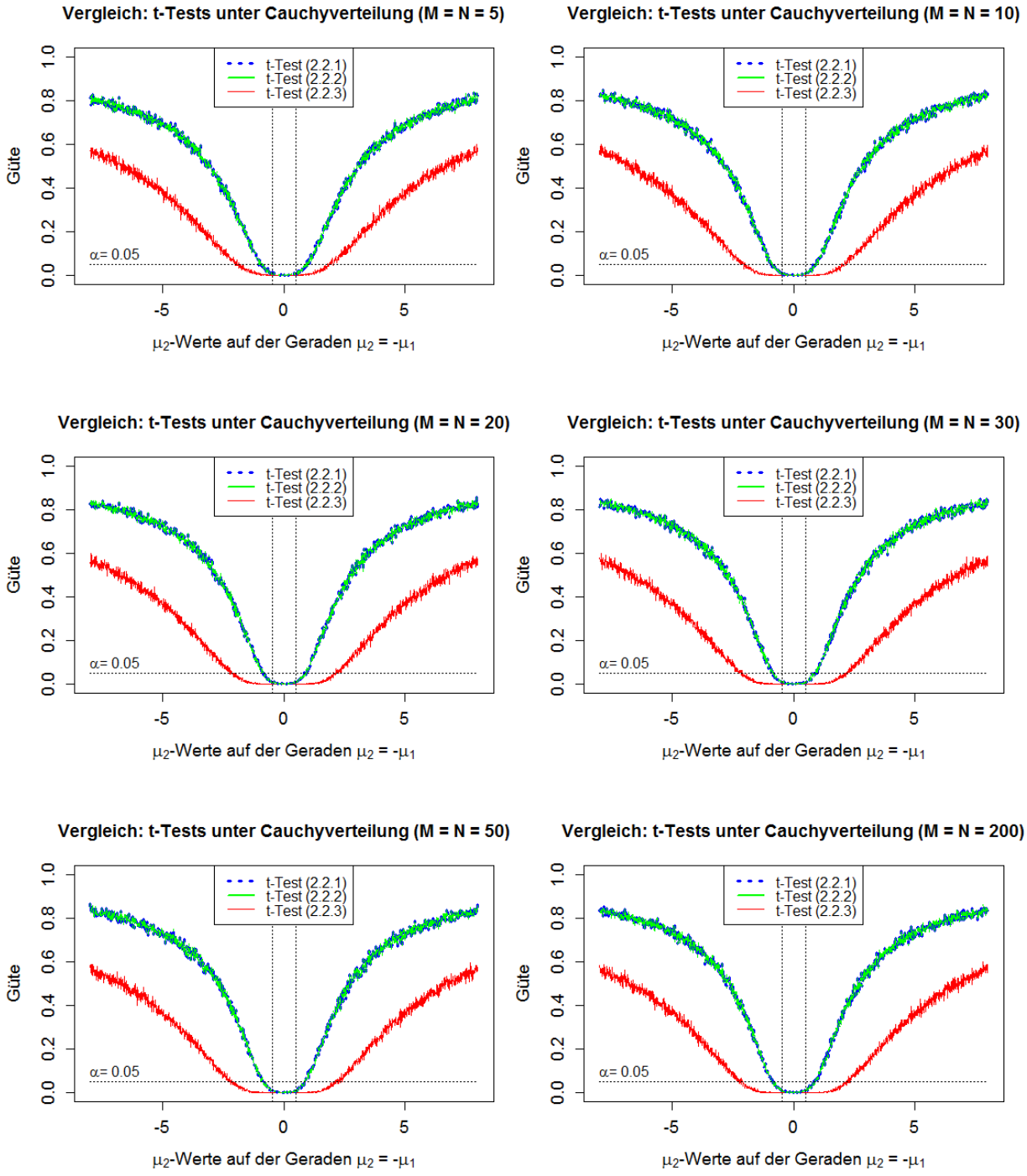


Abbildung 7: Gütefunktionen der t-Tests für verschiedene Stichprobenumfänge

dass sich die Gütefunktionen der t-Tests bei Erhöhungen des Stichprobenumfangs von $M = N = 50$ auf 200 nicht verbessert. Zusammenfassend wird die Problematik

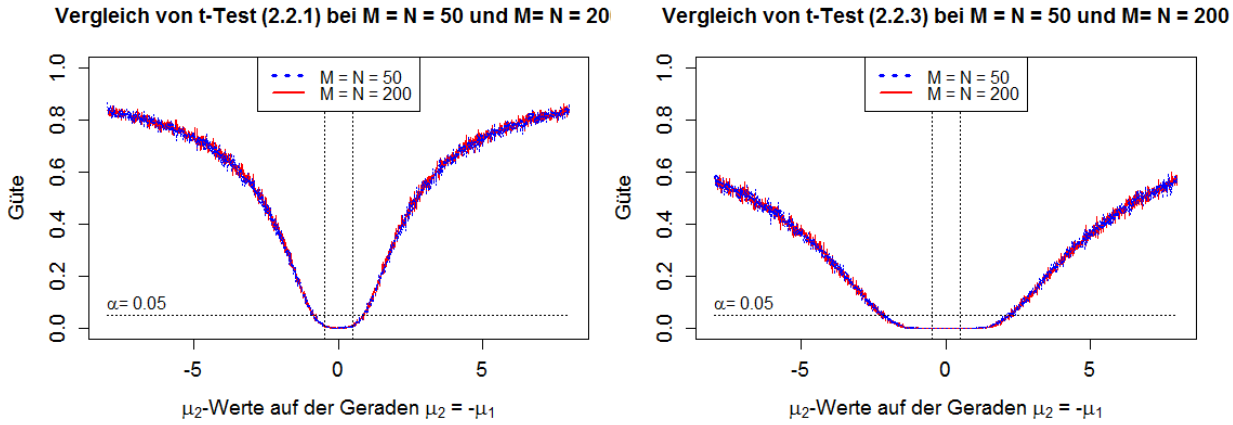


Abbildung 8: Gütefunktionen des t-Tests aus 2.2.1 und 2.2.3 im Vergleich bei unterschiedlichen Stichprobenumfängen unter Cauchyverteilung

bei Verletzung der Modellannahmen bei den t-Tests deutlich.

Der dritte t-Test aus Abschnitt 2.2.3 soll nun für verschiedene Wahlen von σ unter der Cauchyverteilung untersucht werden. Durch erhöhte Wahlen von σ könnte sich Güte des dritten t-Tests verbessern, da der Relevanzparameter δ automatisch kleiner gewählt werden muss und der Nicht-Relevanzbereich verkleinert wird. Dabei wird zum Vergleich der t-Test aus Abschnitt 2.2.2 genommen.

In der Abbildung 9 werden die Gütefunktionen des dritten t-Test für $\sigma = 1, 2, 12$

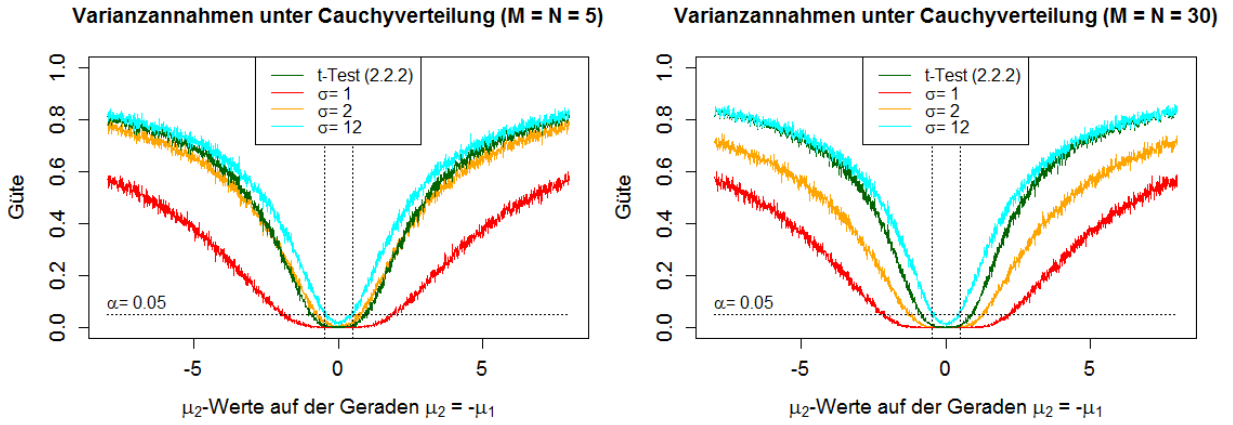


Abbildung 9: Gütefunktionen des t-Tests aus 2.2.3 für erhöhte Varianzannahmen

mit dem zweiten t-Test verglichen. Höhere gewählte Standardabweichungen können durch die kleiner werdenden Nicht-Relevanzbereiche die Güte verbessern. Für eine Wahl von $\sigma = 12$ wird allerdings in der Nähe des Randes der Nullhypothese nicht mehr das Signifikanzniveau eingehalten. Die Ergebnisse des zweiten t-Tests und des dritten t-Test mit $\sigma = 12$ erinnern an die Ergebnisse bei der Normalverteilung. Es liegt für geringe Lageunterschiede eine höhere Güte für den dritten t-Test vor, die aber anschließend von der Güte des zweiten t-Test eingeholt wird.
Da die Wahl der Standardabweichung willkürlich geschieht, da sich diese aus Realisationen cauchyverteilter Zufallsvariablen nicht schätzen lässt und der zweite t-Test ohne diesen willkürlichen Aspekt ähnliche Ergebnisse liefert (ohne die Gefahr das Niveau nicht einzuhalten), sollte dieser Test bevorzugt werden. Dennoch ergeben sich generell in dieser Situation bei den t-Tests nicht zufriedenstellende Resultate.

Spezialfall: $\delta = 0$

In Kapitel 2 wurde gezeigt, dass die Zweistichproben-Relevanz-t-Tests für $\delta = 0$ entartete Fälle des gewöhnlichen Zweistichproben-t-Tests aus Satz 2.4 sind. Die Abbildung 10 bestätigt, dass für $\delta = 0$ keine Unterschiede zwischen den drei Tests unter Normal- oder Cauchyverteilung vorliegen. Hierbei liegen in der linken Grafik

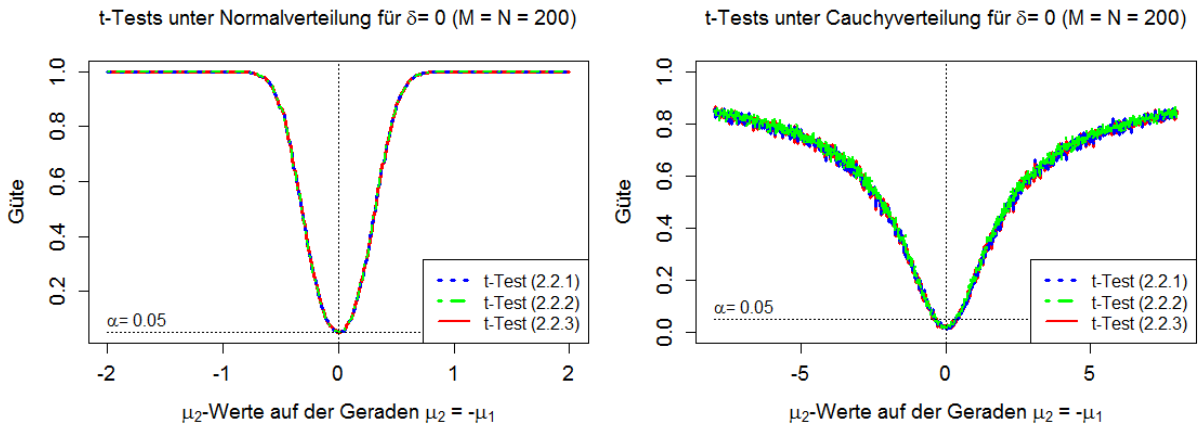


Abbildung 10: Gütefunktionen des t-Tests aus 2.2.3 für erhöhte Varianzannahmen

normalverteilte Stichproben mit $\sigma^2 = 1$ und in der rechten Grafik cauchyverteilte Stichproben mit $\gamma = 1$ vor. Der Annahmebereich besteht lediglich aus dem Punkt $\{0\}$. Je kleiner also δ gewählt wird, desto ähnlichere Ergebnisse liefern die t-Tests.

3.4 Vergleich mit den Tests zur Datentiefe

Die unbefriedigenden Ergebnisse der t-Tests unter Cauchyverteilung sollen nun mit den Relevanz-Tests mit Datentiefe verglichen werden. Zunächst werden die numerische Aspekte, wie die Gitterfeinheit K und Gitterlänge L zu wählen sind, bei den Relevanz-Tests mit Datentiefen untersucht. Anschließend werden die Tests global untereinander verglichen.

3.4.1 Untersuchungen zur geeigneten Berechnung des Supremums

Die folgenden Untersuchungen sollen die Wahl der Gitterfeinheit K und der Gitterlänge L in der nachfolgenden Simulationsstudie rechtfertigen.

Die Wahl der Gitterfeinheit K

Nun soll eine geeignete Skalierung der Feinheit K für eine feste Gitterlänge L = 2 bei der Suche des Supremums im Konfidenzgitter gefunden werden. Dazu wird der Fehler erster Art für $\mu = (1, 0)$ bei einem Stichprobenumfang von $M = N = 30$ zum Relevanzparameter $\delta = 1$ für den Relevanz-Test mit voller Dreier-Tiefe simuliert werden. Dabei werden pro Fehler erster Art 500 Wiederholungen zur Simulation durchgeführt.

Der Wert der Gütefunktion im untersuchten Parameter $\mu = (1, 0)$ entspricht dem Fehler erster Art auf einem Punkt auf dem Rand des Annahmebereichs. Die Untersuchung des Fehlers erster Art ist in dieser Situation aufschlussreich, da er bis auf eventuellen Abweichungen durch die Asymptotik unter dem Signifikanzniveau $\alpha = 0.05$ liegen sollte. Problematiken bei der Suche des Supremums können daher an einem zu hohen Fehler erster Art abgelesen werden, da zu grobe Gitter zu häufiger falschen Ablehnungen der Nullhypothese führen. Insbesondere lässt sich erwarten, dass am Rand der Nullhypothese der Fehler erster Art am höchsten und daher am nächsten beim Signifikanzniveau $\alpha = 0.05$ liegen sollte. Abbildung 11 zeigt die Ergebnisse für verschiedene Feinheiten K. Zusätzlich wird für das Konfidenzgitter einerseits das Stichprobenmittel und andererseits der Median als Schätzung für die Lageparameter μ_1 und μ_2 verwendet. Formel (S) und (M) beschreiben für jeweils eine Stichprobe $X_1, ..., X_M$ die Grundstruktur des Konfidenzintervalls, welches im

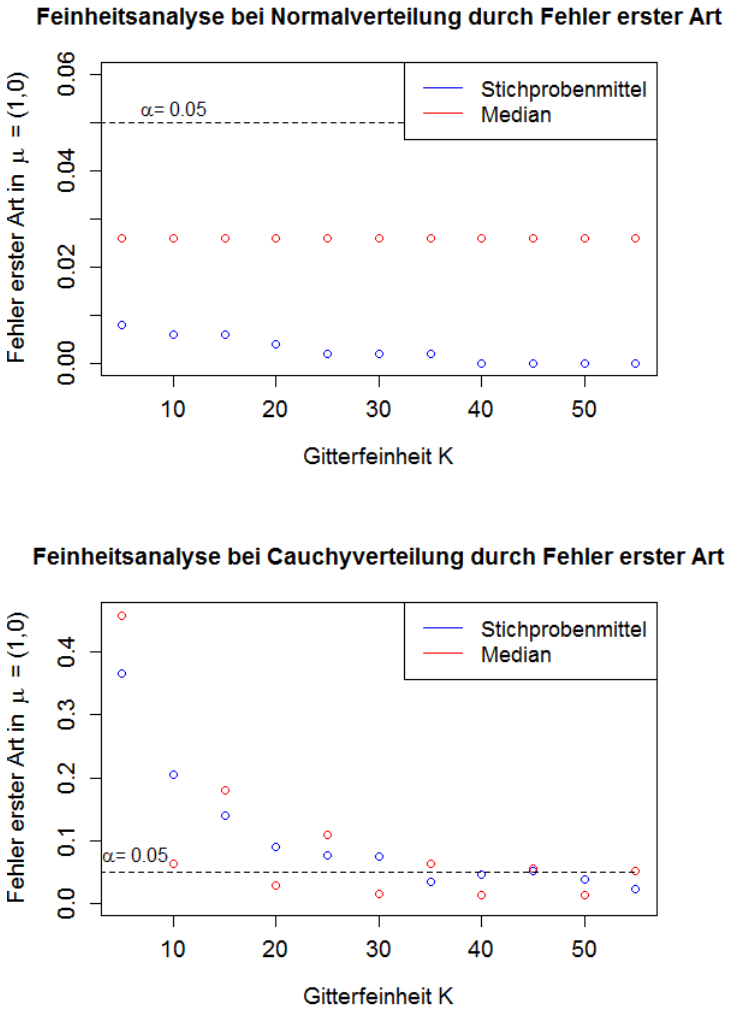


Abbildung 11: Analyse der Gitterfeinheit durch den Fehler erster Art in $\mu = (1, 0)$

Anschluss diskretisiert wird.

$$\left[\overline{X.} - t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X, \overline{X.} + t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X\right] \tag{S}$$

$$\left[\text{med}(X) - t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X, \text{med}(X) + t_{M-1,1-\frac{\alpha}{L}}\widehat{\sigma}_X\right]. \tag{M}$$

Für beide Stichproben werden diese diskretisierten Konfidenzintervalle bestimmt und anschließend wird ein kartesisches Produkt zur Festlegung des zweidimensionalen Gitters verwendet. Links sind in der Abbildung die Fehler erster Art bei normalverteilten und rechts bei cauchyverteilten Stichproben dargestellt.

Unter Normalverteilung wird auch bei kleinen Gitterfeinheiten bei $K \geq 5$ das Si-

gnifikanzniveau $\alpha = 0.05$ sowohl für das Stichprobenmittel als auch den Median eingehalten. Der Fehler erster Art wird für feinere K unter dem Stichprobenmittel kleiner, während er beim Median konstant bleibt. Bei normalverteilten Stichproben wird daher in den nachfolgenden Simulationen bei den Testverfahren mit Tiefen das Stichprobenmittel verwendet. Für die Wahl der Feinheit K gibt die linke Grafik wenige Anhaltspunkte, da die Ergebnisse auch für kleine K stabil sind. Das kann unter anderem daran liegen, dass die Konfidenzgitter das Supremum unter Normalverteilung gut finden und die Konfidenzgitter relativ klein sind. Dennoch wird empfohlen K nicht zu klein zu wählen, um möglichst genaue Simulationsergebnisse zu gewinnen. In dieser Arbeit wird $K = 50$ gewählt.

Unter Cauchyverteilung hält bei Verwendung des Stichprobenmittels erst ab $K \leq 35$ das Signifikanzniveau ein. Allerdings ist keine klare Monotoniestruktur (trotz gleicher Zufallszahlen bei der Simulation) bei erhöhter Feinheit erkennbar. Der Median liefert für Feinheiten $K = 10, 20, 30, 40, 50$ bereits zufriedenstellende Ergebnisse, während für $K = 15, 25, 35, 45$ nicht das Signifikanzniveau eingehalten wird. Man kann zwei Gruppen bei der Monotoniestruktur erkennen. Das kann daran liegen, dass sich das Supremum auf sehr kleinen Regionen finden lässt und man für $K = 15, 25, 35, 45$ Schrittfolgen im Gitter durchläuft, die dieses genau verpassen. Für $K = 10, 20, 30, 40, 50$ läuft die Schrittfolge häufiger nah genug an das Supremum. Ab $K \geq 50$ liefern beide Gruppen zufriedenstellende Ergebnisse, weswegen $K = 50$ als Feinheit empfohlen wird. In der rechten Grafik lassen sich die Ergebnisse für das arithmetische Mittel zwar besser interpretieren. Allerdings ist eine Schätzung für die Lageparameter μ_1 und μ_2 für das arithmetische Mittel nicht konsistent. Da der Median die Lageparameter konsistent schätzt, wird dieser bei der Schätzung des Konfidenzgitters mit Feinheit $K = 50$ verwendet.

Die Wahl der Gitterlänge L

In der nachfolgenden Untersuchung wird das Verhalten des Fehlers erster Art des Relevanz-Tests mit voller Dreier-Tiefe ebenfalls auf einem Randpunkt der Nullhypothese $\mu = (1, 0)$ mit Stichprobenumfang $M = N = 30$ zum Relevanzparameter $\delta = 1$ bei einer Gitterfeinheit von $K = 50$ für verschiedene Gitterlängen L simuliert.

Es werden zur Simulation 500 Wiederholungen durchgeführt. In Tabelle 5 werden die Fehler erster Art unter Normalverteilung dargestellt. Für $L = 1$ wird bereits das

Tabelle 5: Fehler erster Art in $\mu = (1, 0)$ für $K = 50$ und verschiedene L (Normalverteilung)

Gitterlänge L	1	2	3	4	5	6
Fehler erster Art in $\mu = (1, 0)$	0.01	0.002	0.002	0.002	0.002	0.002

Signifikanzniveau eingehalten. Für $L \geq 2$ ist der Fehler erster Art kleiner, bleibt dann aber konstant. Die Werte liegen deutlich unter $\alpha = 0.05$. In Tabelle 6 werden die Fehler erster Art unter Cauchyverteilung aufgeführt. Hier liefert $L = 2$ den nied-

Tabelle 6: Fehler erster Art in $\mu = (1, 0)$ für $K = 50$ und verschiedene L (Cauchyverteilung)

Gitterlänge L	1	2	3	4	5	6
Fehler erster Art in $\mu = (1, 0)$	0.01	0.006	0.01	0.01	0.01	0.012

rigsten Wert, während erhöhte Gitterlängen L den Fehler erster Art erhöhen können. Dies liegt daran, dass die Gitterfeinheit K konstant bleibt und ein zu großes L der Feinheit entgegen wirken kann und das Gitter vergröbert. Dementsprechend ist auf ein geeignetes Verhältnis von K, L zu achten. Für die folgenden Ausführungen wird $L = 2$ gesetzt.

Zusammenfassend sei bei der Wahl von K, L zu beachten, dass vor allem die Gitterfeinheit K hinreichend hoch sein muss, da sonst das Testverfahren nicht mehr das Signifikanzniveau einhalten könnte. Die Wahl der Gitterlänge L spielt eine geringere Rolle, da durch die konsistente Lageschätzung man bereits in der Nähe des Supremums gelangen wird.

3.4.2 Vergleich der Testverfahren unter Normalverteilung

In der Abbildung 12 werden die Tests zur Datentiefe bei normalverteilten E_i für $i = 1, ..., M + N$ mit Varianz $\sigma^2 = 1$ Relevanzparameter $\delta = 1$ verglichen. Die Simulationen sind in 0.01er mit 100 Wiederholungen durchgeführt worden. Die Relevanz-Tests mit Datentiefen halten bei niedrigeren Stichprobenumfängen das Signifikanzniveau $\alpha = 0.05$ ein. Ihre Güte verbessert sich für wachsende Stichprobenumfänge.

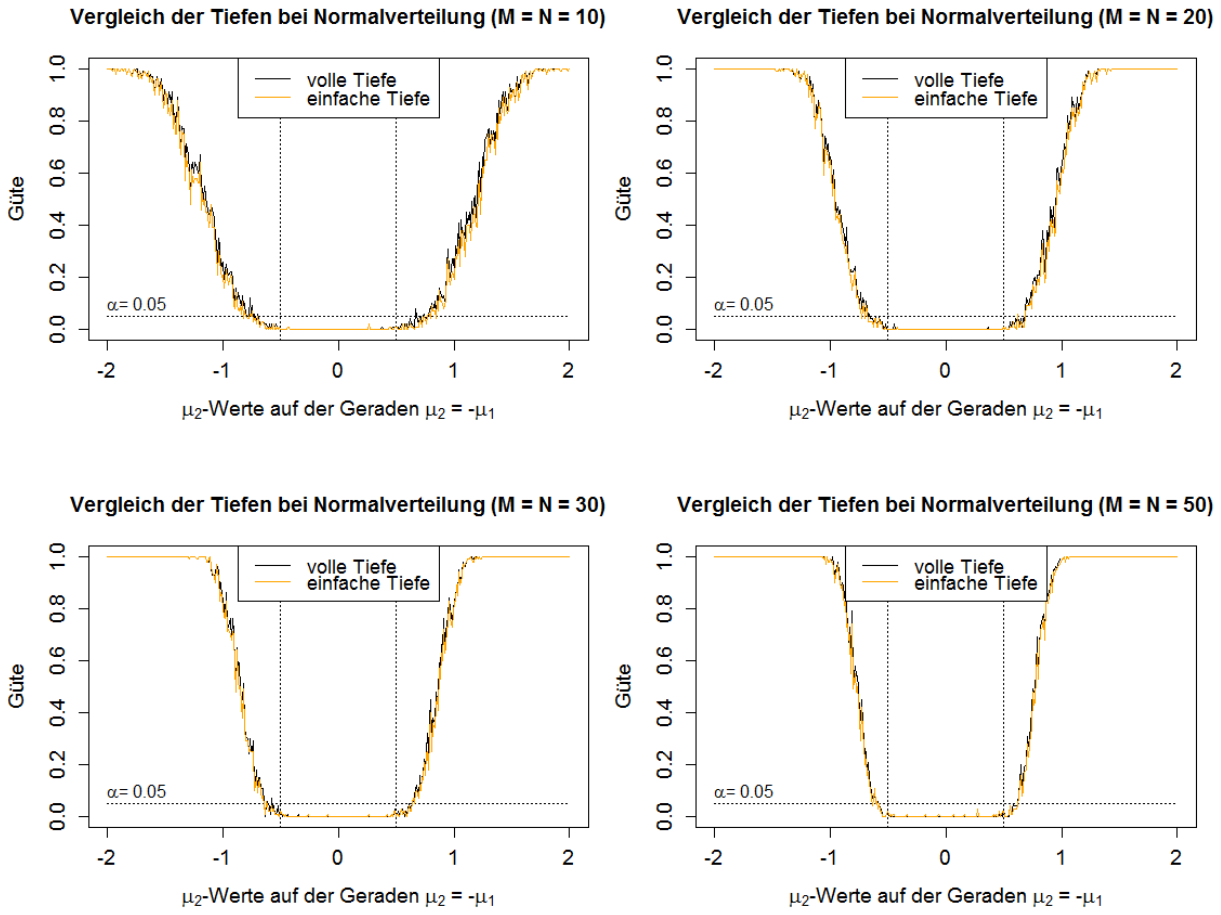


Abbildung 12: Gütefunktionen der Tests zu Datentiefen für verschiedene Stichprobenumfänge (Normalverteilung)

Man kann also auch hier ein *konsistentes* Verhalten vermuten. Die Güteunterschiede zwischen der vollen und vereinfachten Dreier-Tiefe sind gering; die Güte der vollen Dreier-Tiefe ist etwas höher.

In Abbildung 13 werden die t-Test aus Abschnitt 2.2.1 und 2.2.3 mit dem Relevanz-Test zur vollen Datentiefe unter den gleichen Bedingungen bei verschiedenen Stichprobenumfängen verglichen. Die t-Tests sind gleichmäßig deutlich besser als die Relevanz-Test mit Datentiefen. Bei Bekanntheit von normalverteilten Fehlern sollte der t-Tests den Datentiefen vorgezogen werden. Bei höheren Stichprobenumfängen liefern die Datentiefen dennoch relativ zufriedenstellende Ergebnisse.

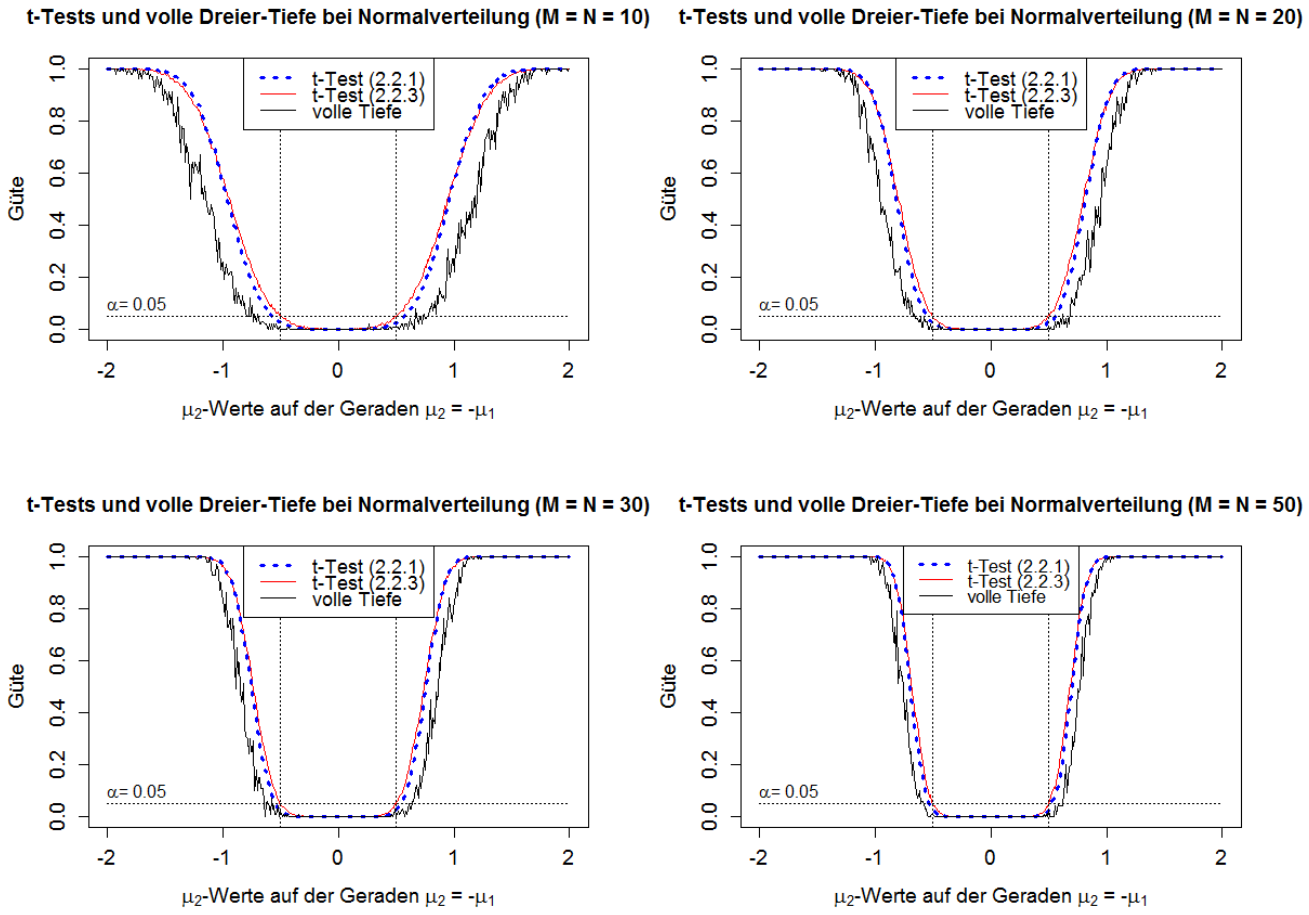


Abbildung 13: Gütefunktionen der t-Tests und der Tests mit Datentiefen für verschiedene Stichprobenumfänge (Normalverteilung)

3.4.3 Vergleich der Testverfahren unter Cauchyverteilung

In der Abbildung 14 werden die Relevanz-Tests zur Datentiefe bei cauchyverteilten E_i für $i = 1, ..., M + N$ mit Skalenparameter $\gamma = 1$ und Relevanzparameter $\delta = 1$ verglichen. Die Simulationen sind in 0.01er mit 100 Wiederholungen durchgeführt worden. Die Relevanz-Tests mit Datentiefen halten auch hier das Signifikanzniveau $\alpha = 0.05$ bei niedrigeren Stichprobenumfängen ein. Die Güte beim Test zur vollen Dreier-Tiefe ist deutlich höher als beim Test mit vereinfachter Dreier-Tiefe. Außerdem erhöht sich die Güte bei wachsenden Stichprobenumfang bei der vollen Dreier-Tiefe intensiv, während sie bei der vereinfachten Dreier-Tiefe nur langsam wächst. Bei der vollen Dreier-Tiefe ist die Vermutung der *Konsistenz* daher naheliegend.

In Abbildung 15 wird der t-Test aus Abschnitt 2.2.1 nun mit den Relevanz-Tests

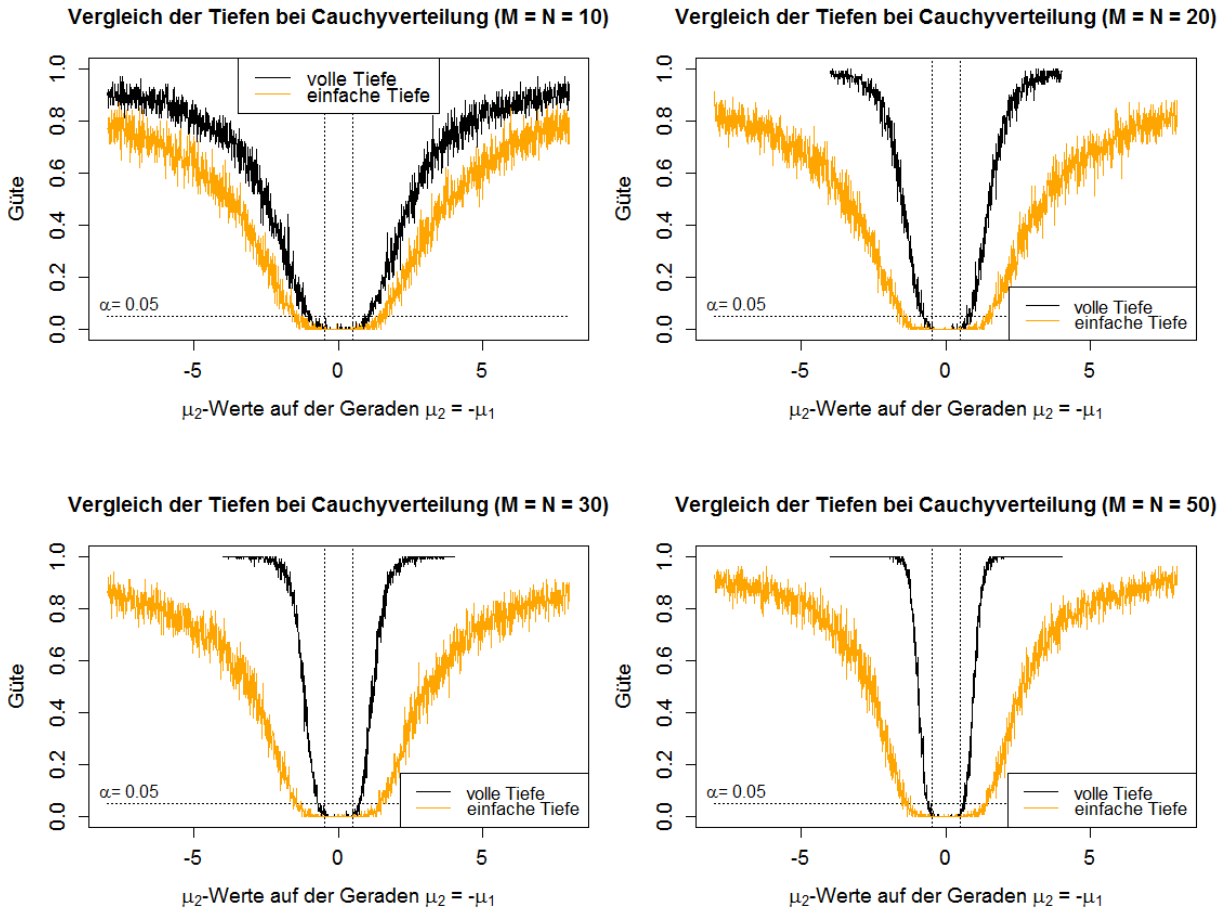


Abbildung 14: Gütefunktionen der Tests zu Datentiefen für verschiedene Stichprobenumfänge (Cauchyverteilung)

zur vollen Dreier-Datentiefe unter den gleichen Bedingungen verglichen. Der t-Test aus Abschnitt 2.2.3 wird in diesem Vergleich wegen der Problematik mit der Wahl einer Standardabweichung σ nicht einbezogen. Auch hier sind die Unterschiede in der Güte beim Test mit voller Dreier-Tiefe und des t-Tests, vor allem bei höheren Stichprobenumfängen, sehr groß. Im Gegensatz zur vollen Dreier-Tiefe kann beim t-Test keine deutliche Verbesserung der Güte bei wachsenden Stichprobenumfang erkannt werden. Bei Bekanntheit von cauchyverteilten Fehlern ist die volle Dreier-Tiefe zu bevorzugen.

In Abbildung 16 wird der Test mit vereinfachter Dreier-Tiefe mit dem t-Test verglichen. Der Test mit vereinfachter Dreier-Tiefe liefert bei geringen Lageunterschieden schlechtere Ergebnisse als der t-Test. Bei höheren Lageunterschieden lässt sich an-

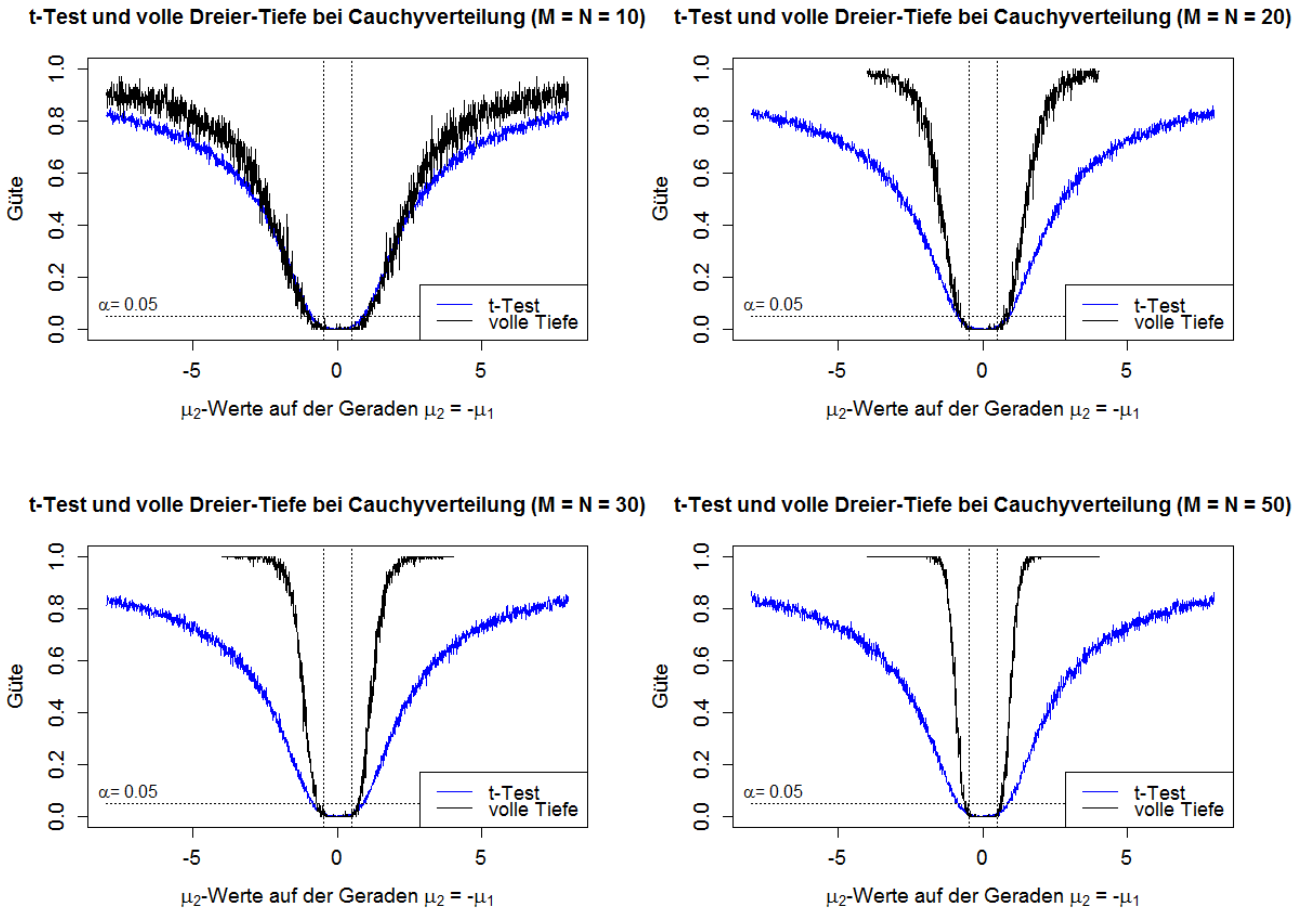


Abbildung 15: Gütefunktionen des t-Tests und der Tests mit voller Datentiefe für verschiedene Stichprobenumfänge (Cauchyverteilung)

nehmen, dass die Güte des t-Test von allen Tests am niedrigsten ist. Insbesondere lässt sich eine Verbesserung der vereinfachten Dreier-Tiefe bei erhöhtem Stichprobenumfang erkennen und die Konsistenz des Tests vermuten.
Insgesamt liefert die volle Dreier-Tiefe stets bessere Resultate als die vereinfachten Dreier-Tiefe und bietet für beide Verteilungsklasse stabile Ergebnisse.

3.5 Untersuchungen der Testverfahren mit Kontaminationen

Zur Untersuchung des Einflusses von Ausreißern werden nun die Gütefunktionen des t-Tests aus Abschnitt 2.2.3 und des Relevanz-Tests beruhend auf der vollen Dreier-Tiefe mit kontaminierten Daten verglichen. Dabei wird lediglich die zweite Stichprobe durch ein Einpunkt-Maß kontaminiert.

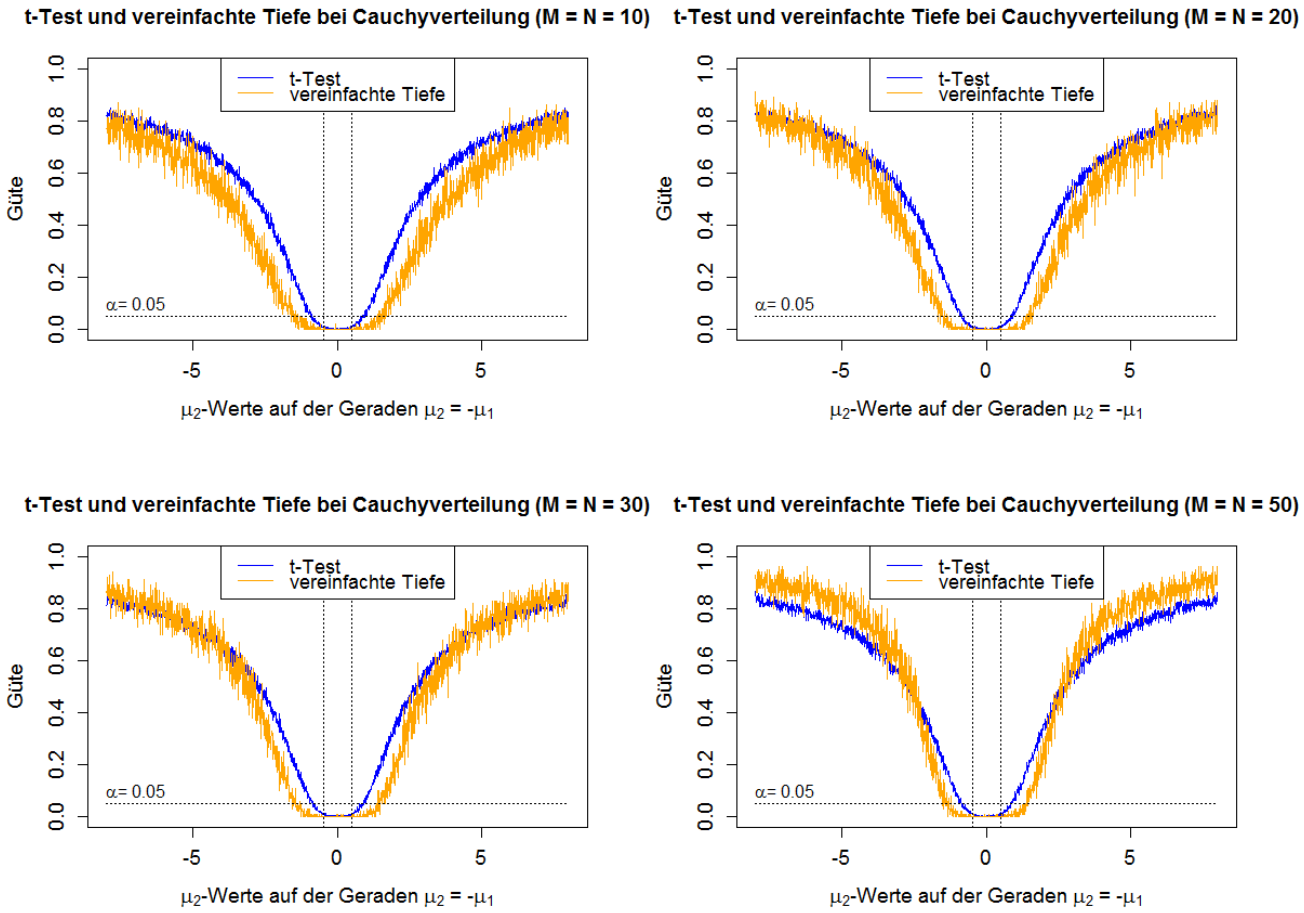


Abbildung 16: Gütefunktionen des t-Tests und der Tests mit vereinfachter Datentiefe für verschiedene Stichprobenumfänge (Cauchyverteilung)

Die exakte Modellierung lautet wie folgt: Seien $X_1, ..., X_M \sim \mathcal{N}(\mu_1, \sigma^2)$ für eine feste Varianz $\sigma^2 \in \mathbb{R}_+$ und einem variablen, unbekannten $\mu_1 \in \mathbb{R}$. Ferner sei die kontaminierte Stichprobe $Y_1, ..., Y_N \sim (\mathcal{N}(\mu_2, \sigma^2))_{\delta_a, \varepsilon}$ (vgl. Abschnitt 2.4.2, S.43ff), wobei δ_a das Einpunkt-Maß im Punkt $a \in \mathbb{R}$ sei, $\varepsilon \in (0, 1)$ die Intensität der Kontamination mit dem Einpunkt-Maß δ_a angibt und $\mu_2 \in \mathbb{R}$ der variable, unbekannte Mittelwert der zweiten Stichprobe ohne Kontamination ist. Die kontaminierte Stichprobe wird dabei durch ein vorgeschaltetes Bernoulli-Experiment simuliert. Mit Wahrscheinlichkeit $(1 - \varepsilon)$ wird eine $\mathcal{N}(\mu_2, \sigma^2)$-Zufallszahl gezogen. Ansonsten ergibt sich die feste Zahl a als Realisation des Einpunkt-Maßes δ_a.

Im Folgenden wird für $\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$ und $a \in \{2, 3\}$ die Gütefunktion des t-Tests aus Abschnitt 2.2.3 und des Relevanz-Tests bezüglich voller Dreier-Tiefe beim

Stichprobenumfang von $M = N = 20$ mit zweidimensionalen Parameterbereich in den Abbildungen 17-20 bzw. 21-24 dargestellt. Die Feinheit des 2D-Gitters sei 0.1 und für jeden Punkt werden 100 Wiederholungen zur Simulation durchgeführt. Die Wiederholungszahl an Simulationen wird wegen der hohen Rechenzeit auf dem 2D-Gitter und bei der Berechnung der Tiefe so niedrig gewählt. Dennoch lassen sich bereits Tendenzen der Gütefunktion erkennen.

Betrachtet man zunächst die Resultate des t-Tests in den Abbildungen 17-20 in der linken Spalte, so wird eine Tendenz zur Rotation des Nicht-Relevanz-Bereichs sichtbar, durch die das Signifikanzniveau für hohe ε unter der Nullhypothese im unteren linken Bereich nicht mehr eingehalten wird. In der rechten Spalte ist ein ähnliches Verhalten beim Relevanz-Test mit voller Dreier-Tiefe erkennbar. Allerdings sind die Auswirkungen auf das Einhalten des Signifikanzniveaus für $\varepsilon \in \{0.05, 0.1, 0.2\}$ geringer als beim t-Test. Für $\varepsilon = 0.3$ besitzt allerdings auch die Gütefunktion des Relevanz-Tests mit voller Dreier-Tiefe, ähnlich wie der t-Test, viele Regionen mit kritisch hohen Werten, sodass das Signifikanzniveau nicht eingehalten wird.

Die Abbildungen 21-24 stellen die gleichen Situationen für δ_3 dar. Der t-Test reagiert hier für gleiche ε sensibler als im Fall für δ_2. Bei dem Relevanz-Test mit Datentiefe sind für $\varepsilon \in \{0.05, 0.1, 0.2\}$ kaum Unterschiede erkennbar. Hier zeigt sich die Robustheit der Datentiefe gegenüber Extremwerten, was anhand der Teststatistik begründet werden kann: Beim t-Test erzielen extreme Realisationen ebenfalls extreme Werte der t-Teststatistik und gegebenenfalls zu erhöhten Ablehnungsraten. Bei der Datentiefe wird das Vorzeichen der Abweichung untersucht, wodurch extreme Werte nicht so ein hohes Gewicht bei der Berechnung der Teststatistik einnehmen. Der Fall $\varepsilon = 0.3$ liefert aber auch bei der Datentiefe nicht mehr akzeptable Werte. Allerdings ist das Resultat beim 20%-Anteil an Kontaminationen zufriedenstellend, womit man hier von einem robusten Verhalten der Tiefe sprechen kann. Zwar wird im Abschnitt 3.5 nur ein Spezialfall untersucht; man könnte andere Formen der Kontaminationen betrachten; allerdings lässt sich erwarten, dass in anderen Untersuchungen mit Kontaminationen ähnliche Ergebnisse erzielt werden.

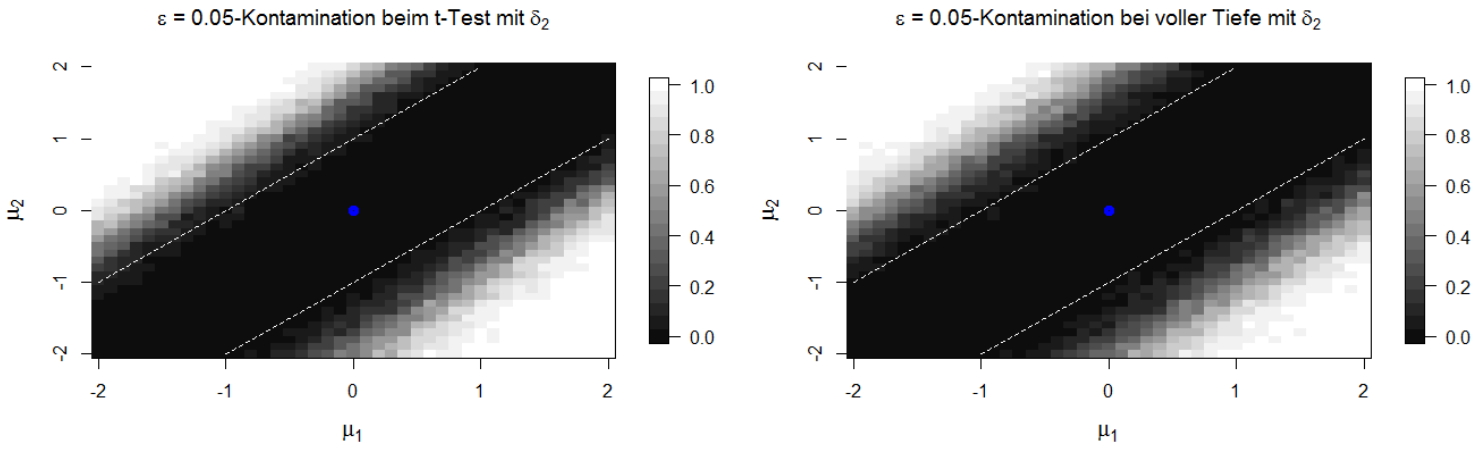


Abbildung 17: Vergleich der Gütefunktionen für $\varepsilon = 0.05$ mit Kontamination durch δ_2 der zweiten Stichprobe

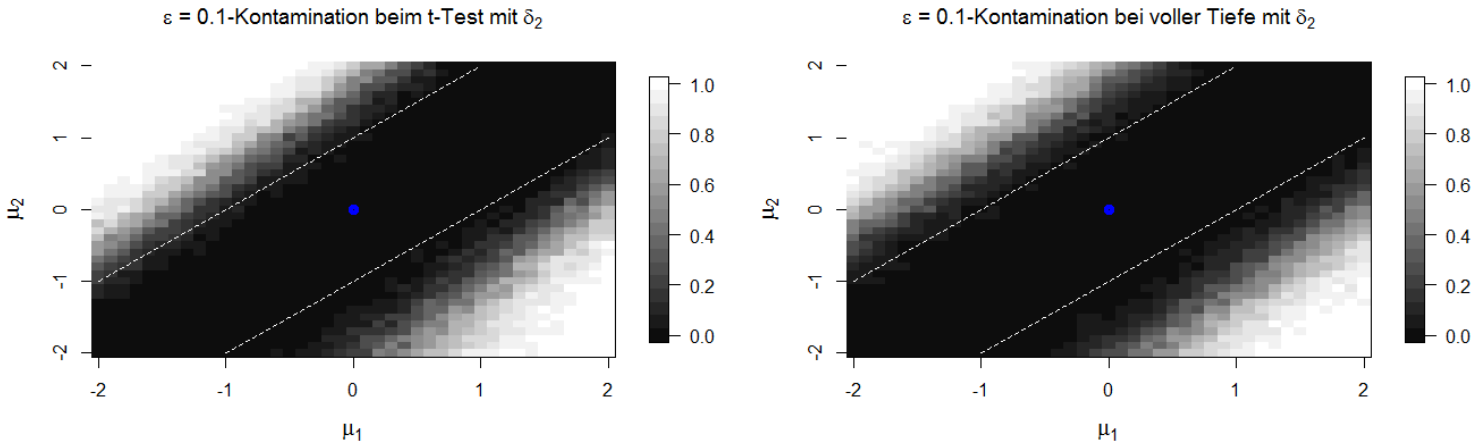


Abbildung 18: Vergleich der Gütefunktionen für $\varepsilon = 0.1$ mit Kontamination durch δ_2 der zweiten Stichprobe

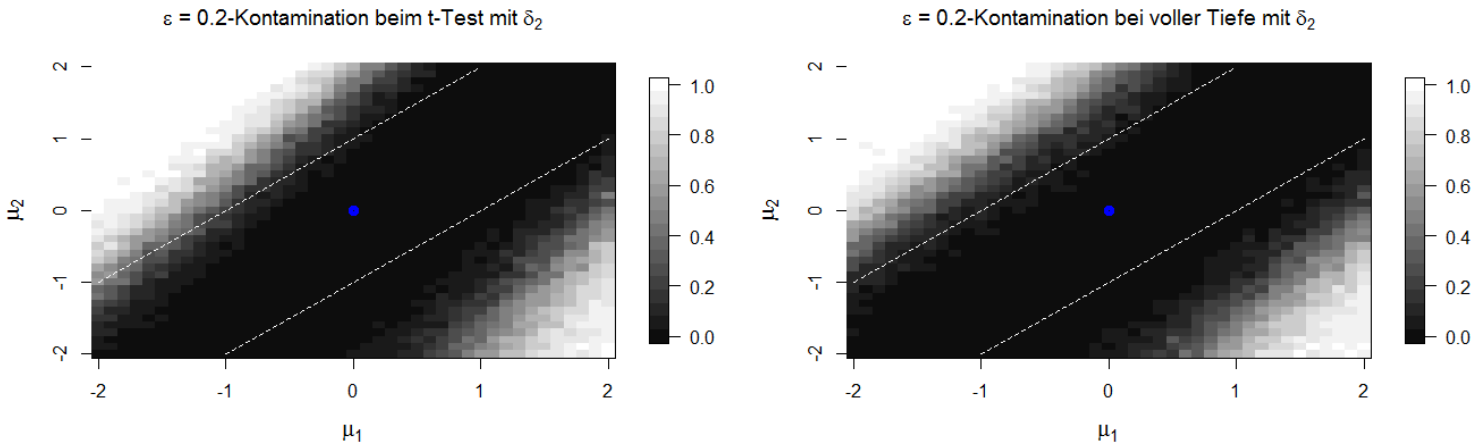
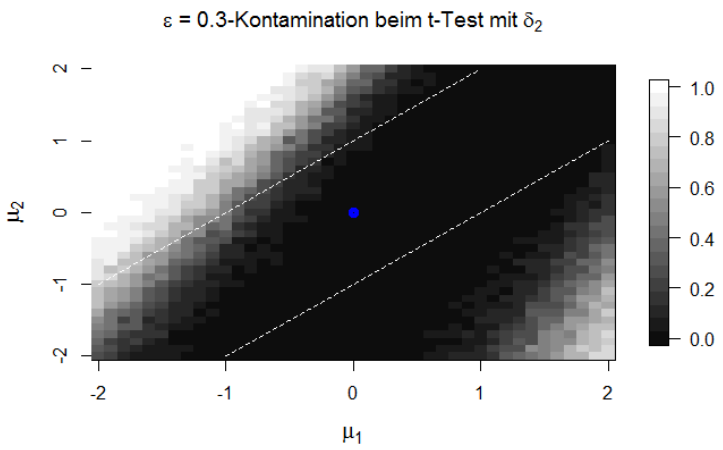


Abbildung 19: Vergleich der Gütefunktionen für $\varepsilon = 0.2$ mit Kontamination durch δ_2 der zweiten Stichprobe



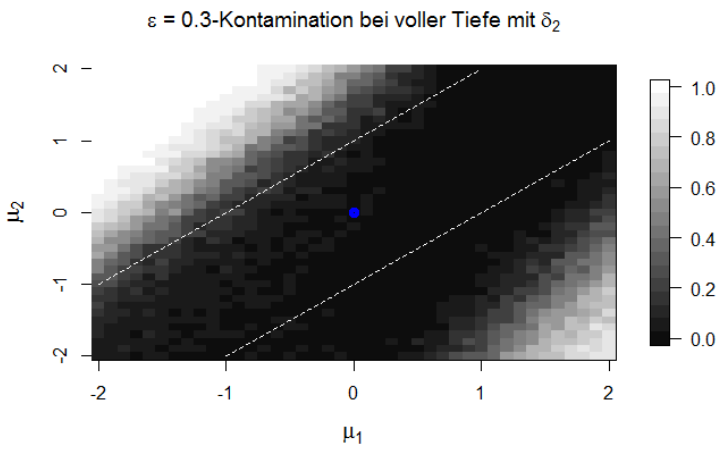
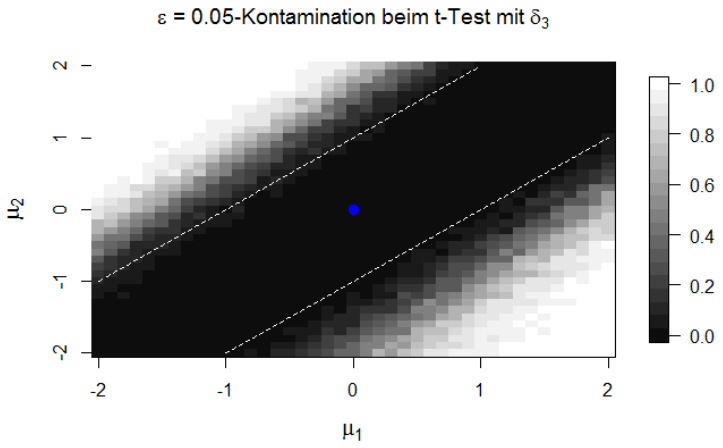


Abbildung 20: Vergleich der Gütefunktionen für $\varepsilon = 0.3$ mit Kontamination durch δ_2 der zweiten Stichprobe



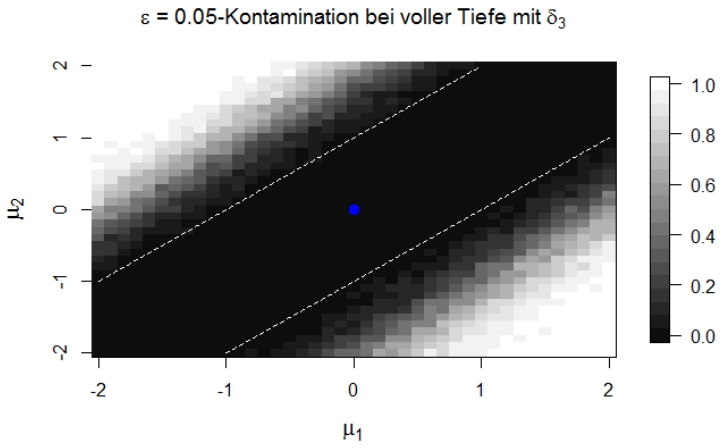
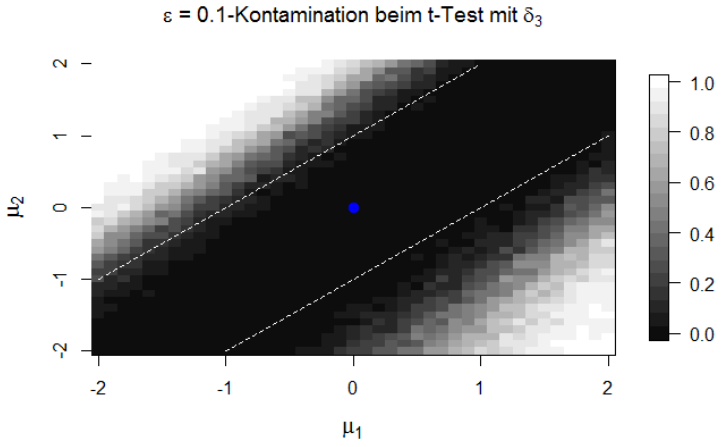


Abbildung 21: Vergleich der Gütefunktionen für $\varepsilon = 0.05$ mit Kontamination durch δ_3 der zweiten Stichprobe



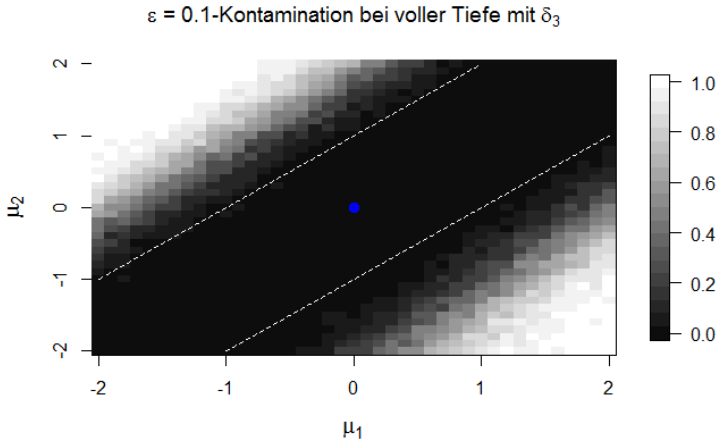
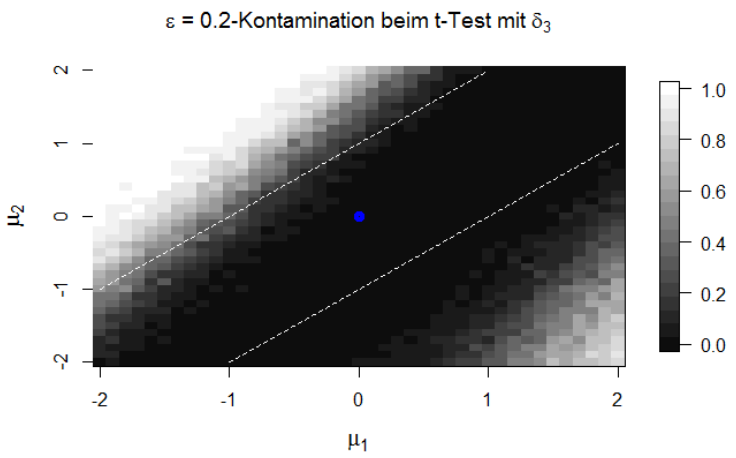


Abbildung 22: Vergleich der Gütefunktionen für $\varepsilon = 0.1$ mit Kontamination durch δ_3 der zweiten Stichprobe



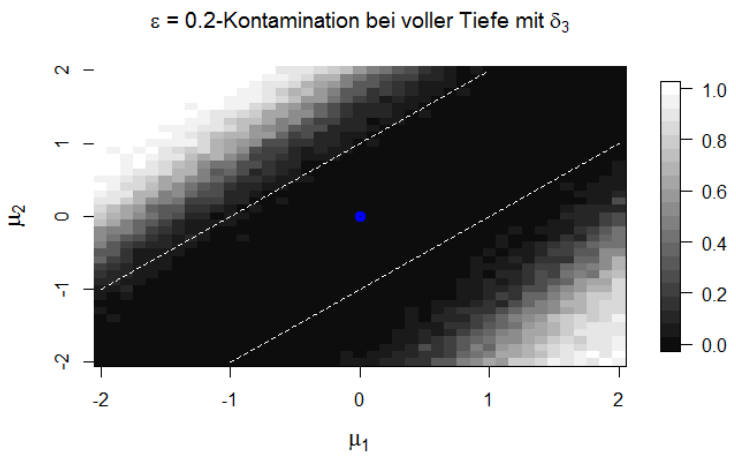
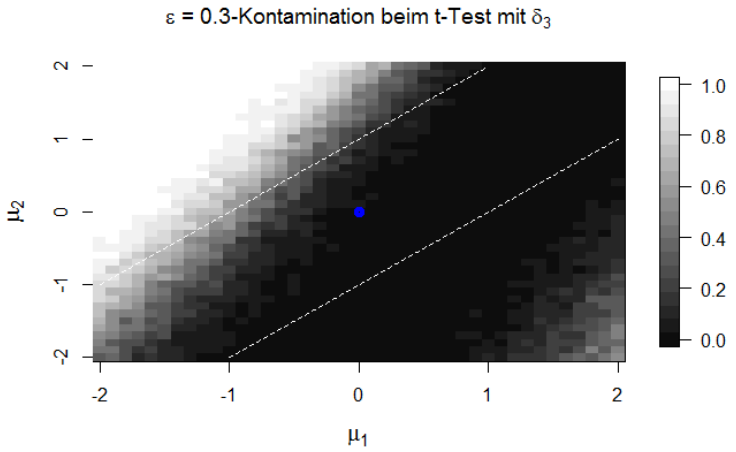


Abbildung 23: Vergleich der Gütefunktionen für $\varepsilon = 0.2$ mit Kontamination durch δ_3 der zweiten Stichprobe



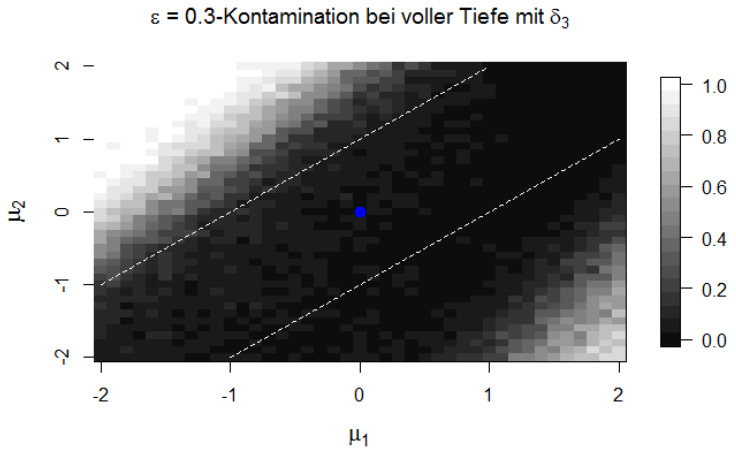


Abbildung 24: Vergleich der Gütefunktionen für $\varepsilon = 0.3$ mit Kontamination durch δ_3 der zweiten Stichprobe

4 Fazit

In diesem Kapitel werden die wichtigsten Resultate dieser Arbeit zusammengefasst und diskutiert. Ferner wird ein Ausblick für weitere Untersuchungsmöglichkeiten anhand von noch offen stehenden Fragestellungen gegeben.

4.1 Zusammenfassung und Diskussion

Beim Vergleich der t-Tests ist bei Bekanntheit der Varianz keiner der gleichmäßig beste. Für geringere Lageunterschiede unter der Alternativen ist der t-Test aus Abschnitt 2.2.3 besser, während für höhere Lageunterschiede die anderen beiden, äquivalenten t-Tests besser sind. Eventuell kann ein gleichmäßig bester t-Test für das untersuche Hypothesenpaar (H) gefunden werden, wenn nur unter den unverfälschten Tests verglichen wird. Lediglich der t-Test aus Abschnitt 2.2.3 ist unverfälscht. Eine solche Optimalitätsaussage existiert für einen Äquivalenztest, der eine analoge Gestalt zu diesem t-Test besitzt (Wellek (2010), S.130).
Zu niedrige Wahlen der Varianzen können beim t-Test aus Abschnitt 2.2.3 zu erheblichen Güteverlusten führen, während bei zu hohen Wahlen der Varianz nicht die Einhaltung des Signifikanzniveaus gewährleistet werden kann. Bei der Verwendung des t-Tests aus Abschnitt 2.2.3 muss daher sichergestellt werden, dass die Varianz der Daten genau bekannt ist. Vor allem muss abgesichert werden, dass die Varianz eher unter- als überschätzt wird, da das Nichteinhalten des Signifikanzniveaus völlig unbefriedigend wäre. Um das Risiko des Nichteinhaltens des Signifikanzniveaus zu vermeiden, wird daher im Allgemeinen von der Verwendung dieses t-Tests abgeraten. Die anderen t-Tests weisen keine derartig gravierende Nachteile auf. Insbesondere liefern alle t-Tests bei hohen Stichprobenumfängen approximativ identische Ergebnisse. Daher sollte nur in Praxissituationen mit sehr geringem Stichprobenumfang bei sehr guter Einschätzung der Varianz der t-Test aus Abschnitt 2.2.3 infrage kommen dürfen.
Bei cauchyverteilten Fehlern ist der Test mit voller Dreier-Tiefe besser als die t-Tests; unter Normalverteilung ist er etwas schlechter. Vor allem in Kontexten, in denen erwartungsgemäß extreme Werte erhoben werden, sollte der Test mit voller

Dreier-Tiefe bevorzugt werden. Falls eine Normalverteilungsannahme nicht gerechtfertigt werden kann, sollte ebenso über die Verwendung des Tests mit voller Dreier-Tiefe nachgedacht werden. Da dieses Testverfahren auch für normalverteilte Fehler zufriedenstellende Ergebnisse liefert, ist generell ein nicht so hoher Güteverlust im Vergleich zum t-Test zu erwarten.
Das Testverfahren mit vereinfachter Dreier-Tiefe ist in allen Untersuchungen schlechter als der Test mit voller Dreier-Tiefe. Daher sollte stets die volle Dreier-Tiefe, trotz höherer Rechenzeit, bevorzugt werden. Ein Grund für die schlechten Ergebnisse der vereinfachten Dreier-Tiefe kann die geringe Untersuchungsfläche zwischen den Stichproben sein. Sie betrachtet lediglich Vorzeichenwechsel der umliegend benachbarten Punkten, was bei zwei Stichproben ungünstig ist, da nicht die Unterschiede zwischen den Stichproben so zur Geltung kommen kann. Eventuell könnten andere Ordnungen der zwei Stichproben untereinander die Effizienz der vereinfachten Tiefe erhöhen.
Die Untersuchungen mit Kontaminationen zeigen in einem Spezialfall, dass die volle Dreier-Tiefe bei 20%-Anteil an Kontaminationen im Gegensatz zum t-Test zufriedenstellende Ergebnisse liefert. Das heißt, dass die volle Dreier-Tiefe ein robustes Verhalten gegenüber extremen Werten aufweist. Man kann anhand der Gestalt der Datentiefe erwarten, dass auch in anderen Ansätzen bei Modellierungen mit Kontaminationen ein ähnliches robustes Verhalten beobachtet werden wird.

4.2 Ausblick

Ideen für weitere Untersuchungen sollen abschließend skizziert werden. Die primäre Ausgangsfrage bleibt, in welchen Situationen sich t-Tests und in welchen Situationen sich der Test mit voller Dreier-Tiefe besser eignen, wenn weder Cauchy- noch Normalverteilung vorliegt. Außerdem ist in Anwendungssituationen in der Regel unbekannt, welcher Verteilung die Daten entsprechen. Mittels Anpassungstests, wie dem Shapiro-Wilk-Test bzw. dem Kolmogorov-Smirnov-Test (Sachs und Hedderich (2015), S.466 bzw. S.461ff), können lediglich Entscheidungen gegen eine Normalverteilung bzw. eine andere feste Verteilung getroffen werden. Wird von einem Anpassungstest der Vorschlag einer Normalverteilung abgelehnt, könnte man anhand der Resultate der Arbeit davon ausgehen, den Relevanz-Test mit voller Dreier-Tiefe zu

verwenden, da der t-Test die Normalverteilung voraussetzt.

Gibt es aber Situationen, in denen die Anpassungstests ihre Nullhypothese (Normalverteilung liegt vor) nicht verwerfen, obwohl der Relevanz-Test mit voller Dreier-Tiefe trotzdem eine deutlich höhere Güte besitzt? Für die Cauchyverteilung wurde in Abschnitt 3.3 der gravierende Güteverlust bei einer falschen Modellannahme demonstriert. Die Cauchyverteilung ist allerdings ein extremes Beispiel und könnte für anwendungsorientierte Fragen zu theoretisch sein. Untersuchungen mit t-verteilten Zufallsvariablen mit unterschiedlichen Freiheitsgraden können dies besser beantworten. Man beachte, dass eine t-Verteilung mit einem Freiheitsgrad der Cauchyverteilung entspricht und t-Verteilungen mit sehr vielen Freiheitsgraden asymptotisch einer Normalverteilung folgen (Büning und Trenkler (1994), S.26). Sie können also als Untersuchungsgegenstand für den Übergang zwischen Cauchy- und Normalverteilung dienen und betrachten nicht lediglich die beiden Extremfälle. Zusätzlich wäre es in diesem Rahmen interessant zu untersuchen, wie Anpassungstests auf t-Verteilungen reagieren und ob die Entscheidung der Anpassungstests zur optimalen Wahl des Relevanz-Tests mit höherer Güte führen würden.

Man kann Untersuchungen vornehmen, in denen man die Gütefunktionen der beiden Typen von Relevanz-Tests für unterschiedliche t-Verteilungen betrachtet. Zusätzlich kann man eine dritte Untersuchungsreihe mit einem vorgeschalteten Anpassungstest vornehmen, der nach einem Entscheidungskriterium (zum Beispiel durch den p-Wert) anhand der Daten angibt, welcher der Relevanz-Tests verwendet wird. Das Entscheidungskriterium muss sich nicht zwingend auf Beibehaltung oder Ablehnung der Nullhypothese für Normalverteilung des Anpassungstests festlegen. Eventuell findet man einen optimalen Schwellenwert (Cutpoint) für den p-Wert für eine Maximierung der Güte. Ferner lassen sich auch andere symmetrische Verteilungsklassen in diesem Rahmen untersuchen.

Außerdem lassen sich weitere Relevanz-Tests mit einem verallgemeinerten Nicht-Relevanzbereich der Form $[\delta_1, \delta_2]$ für $\delta_1 < 0 < \delta_2$ formulieren. In Wellek (2010) wird dies für Äquivalenztests mit Äquivalenzbereich $[\delta_1, \delta_2]$ entwickelt. Umgekehrt können ebenso die Ideen der Relevanz-Tests dieser Arbeit zur Entwicklung von Äquivalenztests beitragen. Beispielsweise kann man einen Äquivalenztest basierend auf

der Tiefe entwickeln.

Ein weiterer wichtiger Aspekt ist bei der Datentiefe eine Verbesserung der Algorithmen für die Suche des Supremums, um die Rechenzeit zu verkürzen. Die Berechnungen in dieser Arbeit erfolgten mit simplen Methoden, ohne die Struktur der vorliegenden Daten genauer zu berücksichtigen. In der Praxis spielt dieser Faktor gerade eine Rolle, wenn der Test mit der Dreier-Tiefe bei großen Datenmengen angewendet werden soll. Außerdem liefert die wiederholte Berechnung von vollen Tiefen, wie hier im Rahmen von Simulationsstudien zur Ermittlung der Güte, sehr hohe Rechenzeiten. In Kustosz, Müller und Leucht (2016) wird eine asymptotische Gleichheit bis auf eine stochastische Nullfolge von der vollen Dreier-Tiefe gefunden. Man kann davon ausgehen, dass diese asymptotische Darstellung eine geringere Rechenzeit hat, da lediglich alle Residuen auf ihre Vorzeichen überprüft werden brauchen. Sofern der Fehler nicht zu groß wäre, könnte man für große Stichprobenumfänge mit dieser Darstellung arbeiten, um die Rechenzeit zu reduzieren.

Schließlich würde die Kenntnis der asymptotischen Verteilung von der vollen $(K+1)$-Tiefe für allgemeine $K \geq 3$ zu weiteren Relevanz-Tests führen. Eventuell liefert eine volle Vierer-Tiefe beispielsweise bessere Testverfahren. Außerdem können Verfahren für Mehrstichproben aufbauend auf der Datentiefe konstruiert und studiert werden.

Literatur

Bauer, H. (2002): *Wahrscheinlichkeitstheorie*, 5. Aufl., Walter de Gruyter, Erlangen.

Büning H. und Trenkler G. (1994): *Nichtparametrische statistische Methoden*, 2. Aufl., Walter de Gruyter, Berlin/Hannover.

Czado, C. und Schmidt, T. (2011): *Mathematische Statistik*, 1. Aufl., Springer, München/Leipzig.

Hoeffding W. and Robbins H. (1948): The Central Limit Theorem for Dependent Random Variables. *Duke Mathematical Journal*, 15, 773-780.

Georgii, H. (2002): *Stochastik - Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 1. Aufl., Walter de Gruyter, München.

Klenke, A. (2006): *Wahrscheinlichkeitstheorie*, 1. Aufl., Springer, Mainz.

Kustosz, C., Leucht, A. and Müller, C. (2016): Tests based on simplicial depth for AR(1) models with explosion. *Journal of Time Series Analysis* 37, 763-784.

Kustosz, C. and Müller, C. (2014). Analysis of crack growth with robust, distributionfree estimators and tests for nonstationary autoregressive processes. *Statistical Papers* 55, 125-140.

Kustosz, C., Müller, C. and Wendler, M. (2016). Simplified simplicial depth for regression and autoregressive growth processes. *Journal of Statistical Planning and Inference* 173, 125-146.

Kustosz, C. and Szugat, S. (2016). *rexpar: Simplicial Depth for Explosive Autoregressive Processes*. R package version 1.1.

Mizera, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* 30, 1681-1736.

Müller, C. (2005): Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis* 95, 153-181.

Müller C. und Denecke L. (2013): *Stochastik in den Ingenieurswissenschaften - Eine Einführung mit R*, 1. Aufl., Springer, Dortmund.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rousseeuw, P.J. and Hubert, M. (1999): Regression depth (with discussion). *J. Amer. Statist. Assoc.* 94, 388-433.

Sachs, L. und Hedderich, J. (2015): *Angewandte Statistik*, 13. Aufl., Springer, Kiel.

Schumacher, M. und Schulgen, G. (2002): *Methodik klinischer Studien*, 1. Aufl., Springer, Berlin.

Wellek, S. (2010): *Testing statistical hypothesis of equivalence and noninferiority*, 2. Aufl., Chapman and Hall/CRC, Heidelberg.

Eidesstattliche Versicherung

_____ _____

Name, Vorname Matr.-Nr.

Ich versichere hiermit an Eides statt, dass ich die vorliegende Bachelorarbeit/Masterarbeit* mit dem Titel

_____

_____

_____

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

_____ _____

Ort, Datum Unterschrift

*Nichtzutreffendes bitte streichen

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz - HG -)

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

_____ _____

Ort, Datum Unterschrift