Ludwig-Maximilians-Universität München
Institut für Statistik
Wintersemester 2013/2014

Bachelorarbeit

Vergleich mehrerer Verfahren für multiples Testen bei der Analyse volatiler organischer Komponenten verschiedener Bakterien und Pilze zur Erregerdifferenzierung

vorgelegt von

Katrin Hummrich

München, 2. Dezember 2013

Betreuerin: Frau Prof. Dr. Anne-Laure Boulesteix

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Problematik des multiplen Testens. Dabei werden die beiden Fehlerraten familiy-wise error rate ($FWER$) und false discovery rate (FDR), sowie insgesamt sieben Adjustierungsverfahren zur Kontrolle dieser Fehlerraten behandelt. Dabei handelt es sich um die $FWER$-kontrollierenden Verfahren von Bonferroni, Holm, Hochberg und die beiden Resampling-Verfahren von Westfall&Young, sowie die beiden FDR-kontrollierenden Verfahren von Benjamini& Hochberg und Benjamini&Yekutieli. Eine praktische Anwendung dieser Verfahren findet anhand eines Datenbeispiels statt, das sich mit der Analyse volatiler organischer Komponenten verschiedener Bakterien und Pilze zur Erregerdifferenzierung befasst. In dieser Arbeit werden lediglich die drei Gruppen gram negative, gram positive und Pilze betrachtet und anhand von Tests auf Lageparameter analysiert, bei welchen Molekülen signifikant unterschiedliche Messwerte zwischen den Gruppen beobachtet werden. Anhand der unterschiedlichen Konservativität der Verfahren kann eine feste Reihenfolge angegeben werden, die auch bei diesem Anwendungsbeispiel zu beobachten ist. Insgesamt sind die $FWER$-kontrollierenden Verfahren konservativer, was bedeutet, dass sie weniger Hypothesen ablehnen. Durch schrittweises Vorgehen bei der Adjustierung der p-Werte kann jedoch ihre Power verbessert werden. Einige Messwerte der insgesamt 200 erfassten Moleküle zeigen einen signifikanten Unterschied zwischen den drei Erregergruppen. Jedoch bedarf es noch weiterführender Analysen um festzustellen zwischen welchen Gruppen genau sich diese Unterschiede befinden und um auch einzelne Erreger desselben Erregertyps voneinander unterscheiden zu können.

1 Einleitung

Ganz allgemein tritt das Problem des multiplen Testens immer dann auf, wenn anhand eines Datenmaterials nicht nur eine sondern mehrere Fragestellungen geklärt werden sollen, also mehrere Nullhypothesen aufgestellt werden. Bei der Suche nach Beispielen für multiple Testprobleme landet man sehr oft im Bereich der Medizin oder Biologie. Es handelt sich dabei häufig um Microarray Studien, also die Genanalyse oder um die Analyse eines neuen Medikaments, dessen Wirksamkeit anhand verschiedener Aspekte gemessen wird. Das sind alles Themen die auch zukünftig von Interesse sein werden, wenn sie nicht sogar noch an Bedeutsamkeit gewinnen werden. Und so wird auch die Frage der Behandlung des multiplen Testproblems in der Wissenschaft immer präsenter.

Es gibt mittlerweile viele verschiedene Verfahren um auch beim multiplen Testen eine Aussage über die Irrtumswahrscheinlichkeit machen zu können. In dieser Arbeit sollen einige ausgewählte $FWER$- und FDR-kontrollierende Methoden vorgestellt und verglichen werden. Als praktisches Anwendungsbeispiel dienen hier Daten volatiler or-

ganischer Komponenten verschiedener Bakterien und Pilze. Der genaue Aufbau dieses Datensatzes, sowie die Fragestellung dahinter werden im ersten Abschnitt dargestellt. Anschließend folgt ein Kapitel zur statistischen Theorie, in dem zum einen das multiple Testproblem und die zwei Fehlerraten genau definiert werden und verschiedene Adjustierungsverfahren vorgestellt werden, und zum anderen zwei mögliche Tests auf Lageparameter beschrieben werden, die zur Beantwortung der Fragestellung benötigt werden. Dann folgt der Vergleich der Adjustierungsverfahren. Dieser erfolgt zunächst rein theoretisch im Kapitel 4 und letztlich im Kapitel 5 auch praktisch anhand der Ergebnisse der Analyse des Datenmaterials. Abschließend erfolgt eine Zusammenfassung der Ergebnisse mit Fazit und ein Ausblick auf mögliche weitere interessante Fragestellungen.

2 Daten und Fragestellung

Die Daten stammen aus einem Experiment, das im Klinikum Großhadern durchgeführt wurde. Inhaltlicher Hintergrund dieses Experiments stellt die Identifikation von Erregern dar. Dazu wurden Nährlösungen (LB = lysogeny broth) angesetzt und mit verschiedenen Erregern versetzt. Diese Erreger können drei Gruppen zugeordnet werden, den gram negativen, den gram positiven und den Pilzen. Folgende Erreger wurden betrachtet und mit diesen Abkürzungen versehen:

gram negativ
PV: Proteus vulgaris
ECL: Enterobacter cloacae
KO: Klebsiella oxytoca
KP: Klebsiella pneumoniae
SM: Serratia marcescens
PA: Pseudomonas aeruginosa
EC: Escherichia coli

gram positiv
SA: Staphyloccus aureus
SE: Staphyloccus epidermidis
EFCL: Enterococcus faecalis
EFCM: Enterococcus faecium

Pilze
CA: Candida albicans
CK: Candida krusei.

Mit Hilfe eines Massenspektrometers wurden die volatilen (lateinisch volatilis = fliegend; flüchtig) organischen Komponenten verschiedener Bakterien oder Pilze gemessen. In den mit Erregern angereicherten, sowie in einigen puren Nährlösungen, wurde also

gemessen welche Moleküle enthalten sind und in welcher Menge. Insgesamt wurden dabei 200 verschiedene Moleküle betrachtet. Diese Messungen wurden zu vier Messzeitpunkten (T0, T1, T2, T3), nach 10, 120, 240 und 360 Minuten, durchgeführt. Die Stichprobengrößen der drei Gruppen sind nicht gleich, da zum einen unterschiedlich viele Erreger je Gruppe betrachtet wurden und zum anderen, je nachdem wie viele Versuche geglückt sind, es einen bis neun Messwerte pro Erreger und Zeitpunkt gibt. Zusätzlich wurden noch drei Variablen erstellt, die den Zeitpunkt und den Erregertyp als character und als factor angeben.
Mit Hilfe dieser Daten soll folgende Fragestellung geklärt werden: Können die drei Erregertypen hinsichtlich der Messwerte der verschiedenen Moleküle voneinander unterschieden werden? Also differenzieren sich die Messwerte der volatilen organischen Komponenten bezüglich der Erregertypen signifikant voneinander?
In der weiteren Arbeit sollen folgende Indizes für alle Formeln gelten:
$j = 1, ..., m$ ist der Index für die Variablen, also hier die 200 Moleküle und somit auch für die dazugehörigen Hypothesen. $i = 1, ...n$ steht für die Beobachtungen, in diesem Fall somit die Erreger. n ist folglich die Stichprobengröße und setzt sich aus n_1, n_2 und n_3 zusammen, den Stichprobengrößen der drei Erregertypen gram negative, gram positive und Pilze. $l = 1, ..., k$ ist der dazugehörige Index für diese Gruppen. Permutationen erhalten den Index $b = 1, ..., B$. s, t, u und z dienen als freie Laufindizes.

3 Theorie

3.1 Tests auf Lageparameter

Um die oben genannten Fragestellungen beantworten zu können bedarf es eines Tests auf Lageparameter. Die Nullhypothese geht dabei immer von Gleichheit aus, die Alternativhypothese von Ungleichheit. Bei einem signifikanten Unterschied wird folglich die Nullhypothese abgelehnt. Da bei diesem Datenbeispiel drei Gruppen miteinander verglichen werden sollen, kommen die Varianzanalyse (ANOVA) als parametrischer Test und der Kruskal-Wallis-Test, auch H-Test genannt, als nicht-parametrischer Test in Frage.
Allgemein liegt im Folgenden diese Datenstruktur vor: es sind k Stichprobengruppen gegeben und insgesamt n Stichprobenelemente, wobei $n = \sum_{l=1}^{k} n_l$ und $n_1, n_2, ..., n_k$ die Umfänge der k Stichproben sind. Dabei wird hier vorerst das Vorgehen nur für eine Variable dargestellt. Auf das vorliegende Datenbeispiel bezogen bedeutet das, dass nicht 200 Moleküle, sondern nur ein Molekül betrachtet wird. Das heißt es wird darauf verzichtet, alles mit einem j zu versehen um anzuzeigen, dass das für das j-te Molekül berechnet wird, was die vielen Indizes übersichtlicher machen soll.

3.1.1 Varianzanalyse

Die Varianzanalyse ist eine Verallgemeinerung des t-Tests und kann zum Vergleich beliebig vieler Erwartungswerte verwendet werden ($k > 2$). Da bei diesem Datenbeispiel nur ein Faktor vorliegt, der Erregertyp, wird die einfaktorielle Varianzanalyse betrachtet. Bei dieser soll untersucht werden, ob die einzelnen Stufen des Faktors eine signifikant unterschiedliche Wirkung auf das interessierende Merkmal, hier der Messwert eines Moleküls, haben. Die Nullhypothese und die zugehörige Alternativhypothese für jedes einzelne Molekül j lauten somit: $H_0^j : \mu_1 = \mu_2 = ... = \mu_k$ und $H_1^j : \mu_s \neq \mu_t$ für mindestens zwei μ_l mit $s, t = 1, ..., k$ und $s \neq t$. (vgl. Fahrmeir u.a. (2012): S. 516-519)
Die richtige Anwendung der Varianzanalyse erfordert drei Voraussetzungen, die die Daten erfüllen sollten: die Varianzhomogenität, das heißt es wird angenommen, dass die Varianzen in den jeweiligen Grundgesamtheiten gleich sind, die Normalverteilungsannahme und die Unabhängigkeit aller Beobachtungen. (vgl. Fahrmeir u.a. (2010): S. 527-528) Auf die Prüfung der ersten beiden Annahmen wird später in diesem Abschnitt eingegangen.

Seien x_{li} die Stichprobenwerte, also der i-te Wert in der l-ten Stichprobe ($1 \leq k; 1 \leq i \leq n_l$). Die Gruppenmittelwerte $\bar{x}_l$ sind dann gegeben durch

$$\bar{x}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{li} \tag{1}$$

und das Gesamtmittel $\bar{x}$ durch

$$\bar{x} = \frac{1}{n} \sum_{l=1}^{k} \sum_{i=1}^{n_l} x_{li} = \frac{1}{n} \sum_{l=1}^{k} n_l \bar{x}_l \ . \tag{2}$$

Zentral bei der Varianzanalyse ist die Streuungszerlegung, bei der sich die Summe der Abweichungsquadrate (SAQ) der Stichprobenwerte um das Gesamtmittel ("Q_{gesamt}") in zwei Teile zerlegen lässt:
1. SAQ der Einzelwerte um die Gruppenmittelwerte, also die Streuung innerhalb der Gruppen ("$Q_{innerhalb}$")
2. SAQ der Gruppenmittelwerte um das Gesamtmittel, also die Streuung zwischen den Gruppen ("$Q_{zwischen}$")

$$Q_{gesamt} = Q_{innerhalb} + Q_{zwischen} \tag{3}$$

$$\sum_{l=1}^{k} \sum_{i=1}^{n_l} (x_{li} - \bar{x})^2 = \sum_{l=1}^{k} \sum_{i=1}^{n_l} (x_{li} - \bar{x}_l)^2 + \sum_{l=1}^{k} n_l (\bar{x}_l - \bar{x})^2 \ . \tag{4}$$

Teilt man die SAQ durch die zugehörigen Freiheitsgrade erhält man die mittleren

Quadrate (MQ). Wobei für die Freiheitsgrade gilt: $(n-1) = (n-k) + (k-1)$. Die mittleren Quadrate sind somit definiert durch:

$$MQ_{zwischen} = s^2_{zwischen} = \frac{1}{k-1} \sum_{l=1}^{k} n_l(\bar{x}_l - \bar{x})^2 \tag{5}$$

und

$$MQ_{innerhalb} = s^2_{innerhalb} = \frac{1}{n-k} \sum_{l=1}^{k} \sum_{i=1}^{n_l} (x_{li} - \bar{x}_l)^2 \ . \tag{6}$$

Kommen die Gruppen aus derselben Grundgesamtheit, sollten die Varianzen, also diese mittleren Quadrate, etwa gleich groß sein. Die Prüfgröße um die Nullhypothese $\mu_1 = \mu_2 = ... = \mu_k$ zu testen berechnet sich folgendermaßen

$$\widehat{F} = \frac{MQ_{zwischen}}{MQ_{innerhalb}} = \frac{\frac{1}{k-1} \sum_{l=1}^{k} n_l(\bar{x}_l - \bar{x})^2}{\frac{1}{n-k} \sum_{l=1}^{k} \sum_{i=1}^{n_l} (x_{li} - \bar{x}_l)^2} = \frac{\frac{1}{k-1} \sum_{l=1}^{k} n_l(\bar{x}_l - \bar{x})^2}{\frac{1}{n-k} \sum_{l=1}^{k} s_l^2(n_l - 1)} \tag{7}$$

und gilt $\widehat{F} > F_{(k-1;n-k;1-\alpha)}$, so wird diese Nullhypothese abgelehnt. Das bedeutet, dass sich mindestens zwei μ_l voneinander unterscheiden. (vgl. Sachs/Hedderich (2009): S. 490-491)

Bevor die Varianzanalyse durchgeführt werden kann müssen die bereits erwähnten Annahmen geprüft werden.
Für die Überprüfung der Normalverteilungsannahme wird der Shapiro-Wilk-Test verwendet. Dieser soll auf das Datenbeispiel bezogen, für jedes Molekül j feststellen, ob diese Stichprobe einer normalverteilten Grundgesamtheit entstammt. Jedoch wird auch hier der Einfachheit halber der Index für das Molekül weggelassen.
Die Nullhypothese dieses Tests geht davon aus, dass die Stichprobe $x_1, x_2, ..., x_n$ aus einer normalverteilten Grundgesamtheit stammt. Ist der p-Wert also nicht signifikant kann von einer Normalverteilung ausgegangen werden. Die Idee der zugehörigen Teststatistik $\widehat{W}$ ist es einen Quotienten aus zwei Schätzungen für die Varianz σ^2 darzustellen. Im Zähler ist die Schätzung der Regressionsgeraden im QQ-Plot und im Nenner die Stichprobenvarianz $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})$. Man erhält somit folgende Formel für $\widehat{W}$

$$\widehat{W} = \frac{b^2}{(n-1)s^2} = \frac{(\sum_{i=1}^{n} a_i x_{r_i})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \ , \tag{8}$$

wobei x_{r_i} die, der aufsteigenden Größe nach sortierten, Beobachtungen sind und $b = \frac{R^2 \widehat{\sigma}}{C}$ mit $R^2 = m^T V^{-1} m$, $C = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}}$ und $\widehat{\sigma} = \frac{m^T V^{-1} x}{m^T V^{-1} m}$. Wobei V die Kovarianzmatrix ist und $m^T = (m_1, ..., m_n)$ die erwarteten Ordnungsstatistiken aus einer

Normalverteilung sind. Außerdem gilt $a^T = (a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}}$ und "(...)a_i sind konstante Werte, die aus den Maßzahlen der Ordnungsstatistik einer normalverteilten Zufallsvariablen abhängig vom Stichprobenumfang n erzeugt oder entsprechenden Tabellen entnommen werden können" (Sachs/Hedderich (2009): S. 398).
Ergibt der Quotient 1 liegen die beiden Schätzungen für die Varianz nahe zusammen und es handelt sich um eine Normalverteilung. Kleine Werte von $\widehat{W}$ sprechen für eine Verletzung der Normalverteilungsannahme. (vgl. Sachs/Hedderich (2009): S. 397-398 und vgl. Shapiro/Wilk (1965): S. 592-593)

Ob Homoskedastizität vorliegt wird mittels des Levene-Tests überprüft. Hier wird die Gleichheit der k Varianzen mittels einer einfachen Varianzanalyse getestet. Dabei müssen die k Stichprobengruppen mindestens 10 Beobachtungen aufweisen. Die Nullhypothese lautet dann $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$, im Gegensatz zur Alternativhypothese $H_1 : \sigma_s^2 \neq \sigma_t^2$ für mindestens zwei σ_l mit $s, t = 1, ..., k$ und $s \neq t$. H_0 wird abgelehnt und somit liegt keine Varianzhomogenität vor, wenn $\widehat{F}$ der Varianzanalyse größer ist als $F_{k-1;n-k;1-\alpha}$. $\widehat{F}$ wird nach einer Transformation $y_{li} = |x_{li} - \tilde{x}_l|$ der Beobachtungen, wobei $\tilde{x}_l$ den Median der l-ten Gruppe darstellt, mit der bereits erwähnten Formel (7) aus dem Abschnitt zur Varianzanalyse berechnet. (vgl. Sachs/Hedderich (2009): S. 489-490)

3.1.2 Kruskal-Wallis-Test oder H-Test

Analog zum Wilcoxon-Mann-Whitney-Test, auch bekannt als U-Test, prüft der Kruskal-Wallis-Test, auch H-Test genannt, ob die k Stichproben aus derselben Grundgesamtheit kommen, ob die k Verteilungsfunktionen also gleich sind. Die Nullhypothese und die dazugehörige Alternativhypothese für das Molekül j lauten: $H_0^j : F_1 = F_2 = ... = F_k$ und $H_1^j : F_s \neq F_t$ für mindestens zwei F_l mit $s, t = 1, ..., k$ und $s \neq t$. Wie im vorhergehenden Abschnitt wird die Vorgehensweise des Kruskal-Wallis-Tests für ein beliebiges Molekül j beschrieben, ohne den Index j jedes Mal hinzuzunehmen. Die Messwerte werden der Größe nach aufsteigend sortiert und ihnen Ränge von 1 bis n zugeordnet. Die Prüfgröße des Kruskal-Wallis-Tests lautet:

$$\widehat{H} = [\frac{12}{n(n+1)}] \cdot [\sum_{l=1}^{k} \frac{R_l^2}{n_l}] - 3(n+1) \ , \tag{9}$$

mit R_l als Summe der Ränge der l-ten Stichprobe. Durch die Beziehung $\sum_{l=1}^{k} R_l = \frac{n(n+1)}{2}$ kann kontrolliert werden, ob die Ränge richtig verteilt wurden. H_0 wird abgelehnt, wenn der errechnete Wert $\widehat{H}$ größer oder gleich dem H-Wert aus der Chi-Quadrat-Tabelle ist mit $P \leq \alpha$.
Haben Werte die gleiche Rangzahl wird dies als Bindung bezeichnet. Sind mehr als

25% aller Messwerte in Bindungen muss $\widehat{H}$ mit folgender Formel korrigiert werden:

$$\widehat{H}_{korr} = \frac{\widehat{H}}{1 - \frac{\sum_{u=1}^{z}(t_u^3 - t_u)}{(n^3 - n)}} \ , \tag{10}$$

wobei t_u die Anzahl der jeweils gleichen Rangplätze in der Bindung u aus allen z Bindungen bezeichnet. Ist der Wert von $\widehat{H}$ bereits signifikant ist es nicht notwendig $\widehat{H}_{korr}$ zu berechnen, da der korrigierte Wert immer größer ist als der nicht korrigierte. (vgl. Sachs/Hedderich (2009): S. 514-515)

3.2 Multiples Testen

"Sollen aufgrund eines Datensatzes mehrere Testprobleme anhand von Signifikanztests überprüft werden, spricht man von einem multiplen Testproblem" (Fahrmeir u.a. (2010): S. 428). Da bei dem vorliegenden Datenbeispiel nicht nur für ein Molekül getestet werden soll, ob sich hinsichtlich ihrer Messwerte die drei Erregertypen voneinander unterscheiden, bedarf es hier auch nicht nur eines Tests auf Lageparameter, sondern 200. Und somit liegt ein multiples Testproblem vor. Mit dieser Problematik, sowie einiger ausgewählter Methoden damit umzugehen, befasst sich dieses Kapitel.

3.2.1 Fehlerraten

Wird ein statistischer Test gemacht geht es darum die Entscheidung zu treffen, ob eine vorher formulierte Nullhypothese abgelehnt oder beibehalten werden soll. Dabei können zwei verschiedene Fehlentscheidungen getroffen werden. Die Nullhypothese wird abgelehnt obwohl sie wahr ist. Dies wird Fehler 1. Art bzw. α-Fehler genannt oder auch als falsch-positives Ergebnis bezeichnet. Die andere mögliche Fehlentscheidung liegt darin die Nullhypothese beizubehalten obwohl sie falsch ist. Dies wird analog Fehler 2. Art bzw. β-Fehler oder auch falsch-negatives Ergebnis genannt. Bei nur einem Test wird die Wahrscheinlichkeit den Fehler 1. Art zu begehen durch das Signifikanzniveau α kontrolliert. Werden mehrere Tests simultan durchgeführt ist es jedoch möglich mehrere α-Fehler zu machen, oder auch dass Fehler unterschiedlicher Art gleichzeitig auftreten, was bei der Konzeption der Fehlerraten zur Kontrolle des Fehlers 1. Art berücksichtigt werden muss. (vgl. Zierer (2013): S. 20-21)
Tabelle 1 zeigt die möglichen Ausgänge eines Signifikanztests.

Nullhypothese	nicht abgelehnt	abgelehnt	
wahr	U	V	m_0
falsch	T	S	m_1
	$m - R$	R	m

Tabelle 1: *Mögliche Testausgänge*

m die Anzahl der getesteten Hypothesen ist bekannt, m_0 und m_1 die Anzahl der wahren und falschen Hypothesen sind unbekannt. R stellt eine beobachtete Zufallsvariable dar und S, T, U und V sind nicht beobachtbare Zufallsvariablen. Ziel beim Testen ist es V, die Anzahl der Fehler 1. Art und T, die Anzahl der Fehler 2. Art, möglichst gering zu halten. Bei einem Test mit einer geringen Anzahl von Fehlern 2. Art, spricht man auch von einer hohen Power. (vgl. Dudoit u.a. (2003): S. 73)
Im Folgenden werden zwei Ansätze von Fehlerraten, die helfen sollen den Fehler 1. Art auch bei multiplen Tests zu kontrollieren, vorgestellt.

FWER

Die "family-wise error rate" ($FWER$) ist als die Wahrscheinlichkeit, dass mindestens ein Fehler 1. Art gemacht wird, definiert

$$FWER = \mathbb{P}(V \geq 1) \ . \tag{11}$$

FDR

Die "false discovery rate" (FDR) stellt den erwarteten Anteil von Fehlern 1. Art unter allen abgelehnten Hypothesen dar

$$FDR = \mathbb{E}(Q) \quad \text{mit } Q = \begin{cases} V/R & \text{wenn } R > 0 \\ 0 & \text{wenn } R = 0 \ . \end{cases} \tag{12}$$

Oder anders dargestellt: $FDR = \mathbb{E}(V/R|R > 0)\mathbb{P}(R > 0)$. (vgl. Dudoit u.a. (2003): S. 73)

Ein multipler Test gilt also als kontrolliert zum Niveau α hinsichtlich einer dieser Fehlerraten, wenn gilt $FWER \leq \alpha$ bzw. $FDR \leq \alpha$. Man unterscheidet hierbei zwischen schwacher und starker Korntrolle. Letztere kontrolliert die Fehlerrate unabhängig der Kombination aus wahren und falschen Nullhypothesen. Die schwache Kontrolle hingegen kontrolliert unter der globalen Nullhypothese $H_0^C = \bigcap_{j=1}^{m} H_j$ mit $m_0 = m$, das heißt dass alle Nullhypothesen wahr sind. (vgl. Dudoit (2003): S. 73-74)
Die $FWER$ hat ein sehr strenges Kriterium, die Wahrscheinlichkeit, dass mindestens ein Fehler 1. Art auftritt. Somit stellt sie das konservativere Konzept dar. Dies bietet jedoch im Vergleich zur FDR den Vorteil, dass nicht nur ein Erwartungswert kontrolliert wird. "Allerdings wird hier die Anzahl der abgelehnten Hypothesen und damit indirekt der Anteil wahrer Nullhypothesen mit einbezogen. Damit ist die FDR trotz ihrer Schwächen eine weniger restriktive Alternative zur $FWER$ (...)" (Zierer (2013): S.35). Es gilt $FWER \leq FDR$, wobei Gleichheit eintritt, wenn alle Nullhypothesen wahr sind. Denn in diesem Fall entspricht die Anzahl der fälschlich abgelehnten Hypothesen der Anzahl aller abgelehnten Hypothesen, also $V = R$. Folglich

nimmt Q den Wert 1 für $V > 0$ und den Wert 0 für $V = 0$ an. Damit erhält man $FDR = \mathbb{E}(Q) = 1 \cdot \mathbb{P}(V > 0) = FWER$. Daraus ergibt sich, dass eine Kontrolle der FDR auch eine schwache Kontrolle der $FWER$ gewährleistet. (vgl. Zierer (2013): S. 32/S. 35 und vgl. Dudoit u.a. (2003): S. 74) $FWER$ und FDR sind nahezu gleich, wenn die Anzahl falscher Hypothesen klein ist, und FDR wird umso kleiner als $FWER$ ausfallen je größer die Anzahl falscher Hypothesen ist. (vgl. Sachs/Hedderich (2009): S. 498)

3.2.2 Adjustierungsverfahren

Der Fehler 1. Art wird mit Hilfe von α reguliert. Ist der p-Wert kleiner als das vorgegebene Signifikanzniveau α, bedeutet das bei nur einem Test, dass die Nullhypothese mit einer Irrtumswahrscheinlichkeit von α abgelehnt werden kann. (vgl. Sachs/Hadderich (2009): S. 361) Werden mehrere Tests gemacht steigt die Wahrscheinlichkeit mindestens einen Fehler 1. Art zu machen. Bei m unabhängigen Tests zum Niveau α gilt für die Wahrscheinlichkeit mindestens ein falsch positives Ergebnis zu erhalten: $\alpha^* = 1-(1-\alpha)^m$. Zur Verdeutlichung dieser Problematik ein Beispiel. Sei $\alpha = 0.05$, so ergibt sich für α^* bei m Hypothesen Folgendes in Tabelle 2. (vgl. Fahrmeir u.a.(2010): S. 428)

m	α^*
3	0.143
5	0.226
10	0.401
100	0.994(!)

Tabelle 2: *α^* im multiplen Fall im Verhältnis zur Anzahl m der Hypothesen und dem Signifikanzniveau $\alpha = 0.05$ der einzelnen Tests.*

Um zu verhindern, dass die vorgegebene Fehlerwahrscheinlichkeit überschritten wird, gibt es verschiedene Korrekturverfahren. Generell geht es darum die einzelnen Hypothesen nur dann abzulehnen, falls der zugehörige adjustierte p-Wert kleiner gleich dem vorgegebenen Signifikanzniveau ist. Dabei gibt es verschiedene Vorgehensweisen. Man unterscheidet zwischen single-step Verfahren und schrittweisen Verfahren.
Bei den single-step Verfahren wird die entsprechende Adjustierung für alle Hypothesen gleich und unabhängig der Testergebnisse der anderen Hypothesen durchgeführt.
Die schrittweisen Prozeduren betrachten und adjustieren die Hypothesen nacheinander, sodass vorherige Tests die nachkommenden Ergebnisse beeinflussen. Hierfür werden die noch nicht adjustierten p-Werte der Größe nach sortiert.
Step-down Verfahren beginnen mit dem Adjustieren bei den Hypothesen mit den kleinsten p-Werten, also mit den signifikantesten Hypothesen. Sobald eine Nullhypothese nicht abgelehnt werden kann wird keine weitere Hypothese mehr abgelehnt.
Step-up Prozeduren verfahren umgekehrt. Sie beginnen mit den am wenigsten sig-

nifikanten Hypothesen und sobald eine abgelehnt wurde werden alle folgenden Hypothesen auch abgelehnt. (vgl. Dudoit u.a. (2003): S.78)

Im Folgenden werden einige Adjustierungsverfahren vorgestellt. Zuerst die $FWER$- und anschließend die FDR-kontrollierenden Prozeduren.

Bonferroni

Das Adjustierungsverfahren nach Bonferroni ist eine single-step Prozedur zur Kontrolle der $FWER$ zum Niveau α. Dabei werden alle Hypothesen H_j abgelehnt, deren nicht adjustierter p-Wert kleiner oder gleich $\frac{\alpha}{m}$ ist. Oder anders ausgedrückt, die adjustierten p-Werte nach Bonferroni sind definiert als

$$\tilde{p}_j = min(mp_j, 1) \ . \tag{13}$$

Erklärt wird diese Adjustierung durch folgende Ungleichung, bei der angenommen wird, dass H_j die wahren Nullhypothesen sind mit $j = 1, ..., m_0$:

$$FWER = \mathbb{P}(V \geq 1) = \mathbb{P}(\bigcup_{j=1}^{m_0}(\tilde{P}_j \leq \alpha)) \leq \sum_{j=1}^{m_0}\mathbb{P}(\tilde{P}_j \leq \alpha) \leq \sum_{j=1}^{m_0}\mathbb{P}(P_j \leq \frac{\alpha}{m}) \leq \frac{m_0\alpha}{m} \ ,$$

wobei die letzte Ungleichung aus der Beziehung $\mathbb{P}(P_j \leq x|H_j) \leq x$ für alle x$\in$[0,1] hergeleitet wird und $\tilde{P}_j$ und P_j die Zufallsvariable der adjustierten bzw. nicht-adjustierten p-Werte bezeichnet. (vgl. Dudoit u.a. (2003): S. 78 und vgl. Sachs/Hedderich (2009): S. 498-499)

Da es sich bei den folgenden vier Verfahren um Resampling-Verfahren handelt, soll allgemein das Prinzip dieser Vorgehensweise vorab kurz beschrieben werden. Mit Hilfe von Resampling-Verfahren können p-Werte bestimmt werden, ohne dass unter der Nullhypothese eine Verteilungsannahme gemacht werden muss. Da die empirische Verteilung aus der Stichprobe als Schätzer für die wahre Verteilung dient, kann so die Verteilung indirekt einbezogen werden. Basis von Resampling-Verfahren ist die wiederholte Verwendung einer einmal erhobenen Stichprobe. Hier soll das mit Hilfe von Permutationen geschehen, das heißt durch das mehrmalige Neusortieren der Originalstichprobe. Im multiplen Fall möchte man die Abhängigkeitsstruktur der Teststatistiken erhalten und lässt deshalb dabei jeweils gesamte Beobachtungsvektoren zusammen. Dadurch kann die gemeinsame Verteilung berücksichtigt werden, ohne dass diese explizit bekannt sein muss. Bei der Permutation werden jeweils n_l Beobachtungen zufällig der l-ten Gruppe zugeordnet und die gewünschte Prüfgröße bestimmt. Da die Anzahl möglicher Permutationen oft sehr groß ist, wird meist nur eine Stichprobe aus allen möglichen Permutationen verwendet. (vgl. Zierer (2013): S. 42-43)

single-step minP Prozedur

Das erste Resampling-Verfahren ist die single-step Variante des minP-Verfahrens von Westfall&Young (1993). Allgemein sind die adjustierten p-Werte wie folgt definiert:

$$\tilde{p}_j = \mathbb{P}(\min_{1 \leq s \leq m} P_s \leq p_j | H_0^C) \ , \tag{14}$$

wobei P_s die Zufallsvariable des nicht-adjustierten p-Wertes der s-ten Hypothese bezeichnet und H_0^C die bereits definierte globale Nullhypothese. (vgl. Dudoit u.a. (2003): S. 78)
Dieses Vorgehen lässt sich auch schrittweise darstellen. Für die b-te Resampling-Stichprobe, $b = 1, ..., B$, wird dabei folgendermaßen vorgegangen: Im ersten Schritt wird ein Vektor von nicht-adjustierten p-Werten $p^{*b} = (p_1^{*b}, ..., p_m^{*b})$ für die Nullhypothesen $H_0^i, i = 1, ..., m$, erzeugt. Um (approximativ) die gleiche Verteilung wie die originalen p-Werte unter der globalen Nullhypothese H_0^C zu erhalten kann als Resampling-Verfahren das eben vorgestellte Permutationsverfahren angewendet werden. Für diese Permutation werden dann die nicht-adjustierten p-Werte genau wie bei der Originalstichprobe berechnet. Im zweiten Schritt wird das Minimum der p-Werte der b-ten Resampling-Stichprobe $p_{min}^{*b} = \min_{j=1,...,m} p_j^{*b}$ berechnet.
Schließlich sind die adjustierten p-Werte folgendermaßen definiert:

$$\tilde{p}_j = \frac{\sum_{b=1}^{B} \mathbb{1}(p_{min}^{*b} \leq p_j)}{B} \tag{15}$$

mit $j = 1, ..., m$ und $\mathbb{1}$ als Indikatorfunktion. (vgl. Zierer (2013): S. 46-47)

single-step maxT Prozedur

Alternativ können statt der nicht-adjustierten p-Werte auch die Teststatistiken verwendet werden, wie bei der maxT Prozedur, die zunächst ebenfalls als single-step Variante dargestellt ist (vgl. Dudoit u.a. (2003): S. 78):

$$\tilde{p}_j = \mathbb{P}(\max_{1 \leq s \leq m} |T_s| \geq |t_j| | H_0^C) \ . \tag{16}$$

Das schrittweise Vorgehen ist ebenfalls analog zur minP Prozedur: Im ersten Schritt werden die Teststatistiken $t_1^{*b}, ..., t_m^{*b}$ der b-ten Permutation für jede Hypothese H_0^j berechnet. Im zweiten Schritt wird für eine zweiseitige Alternative an Stelle des Minimums der p-Werte das Maximum der Beträge der Teststatistiken genommen $t_{max_{|.|}}^{*b} = \max_{j=1,...,m} |t_j^{*b}|$ mit $b = 1, ..., B$. Und die adjustierten p-Werte somit folgendermaßen berech-

net:

$$\tilde{p}_j = \frac{\sum_{b=1}^{B} \mathbb{1}(t^{*b}_{max_{|\cdot|}} \geq |t_j|)}{B} \ . \tag{17}$$

(vgl. Zierer (2013): S. 47)

Die single-step Prozeduren sind eher konservativ und haben somit eine geringere Power, was durch ein schrittweises Verfahren deutlich verbessert werden kann (vgl. Dudoit u.a. (2003): S.79). Seien im Folgenden $p_{r_1} \leq p_{r_2} \leq ... \leq p_{r_m}$ die beobachteten geordneten nicht-adjustierten p-Werte und $H_{r_1}, H_{r_2}, ..., H_{r_m}$ die zugehörigen Nullhypothesen.

step-down minP Prozedur

Analog zur single-step minP-Prozedur haben Westfall&Young (1993) auch eine step-down minP-Prozedur mit den adjustierten p-Werten

$$\tilde{p}_{r_j} = \underset{t=1,...,j}{max}\{P(\underset{s\in[r_t,...,r_m]}{min} P_s \leq p_{r_t}|H_0^C)\} \tag{18}$$

entwickelt (vgl. Dudoit u.a. (2003): S. 79-80). Diese verfährt mit der b-ten Resampling-Stichprobe, $b = 1, ..., B$, wie folgt: Auch hier geht es im ersten Schritt um die Berechnung der nicht-adjustierten p-Werte der b-ten Permutation $p_1^{*b}, ..., p_m^{*b}$ für jede Hypothese H_0^j mit $j = 1, ..., m$. Im zweiten Schritt werden zunächst die sukzessiven Minima der nicht-adjustierten p-Werte berechnet, $q_m^{*b} = p_{r_m}^{*b}$ und $q_j^{*b} = min(q_{j+1}^{*b}, p_{r_j}^{*b}), j = 1, ..., m-1$, wobei der Rang r_j nach den beobachteten p-Werten vergeben wird, sodass sich die oben genannte Monotonie der p_{r_j} ergibt. Dabei ist es nicht zwingend, dass die p-Werte $p_{r_j}^{*b}$ der Resampling-Stichprobe dieselbe Monotonie aufweisen wie die p-Werte, die auf der ursprünglichen Stichprobe basieren. Die adjustierten p-Werte werden dann mittels

$$\tilde{q}_{r_j} = \frac{\sum_{b=1}^{B} \mathbb{1}(q_j^{*b} \leq p_{r_j})}{B} \tag{19}$$

berechnet, mit $\mathbb{1}$ als Indikatorfunktion.
Anhand der sukzessiven Maxima

$$\tilde{p}_{r_1} = \tilde{q}_{r_1}, \ \tilde{p}_{r_j} = max(\tilde{q}_{r_j}, \tilde{p}_{r_{j-1}}) \tag{20}$$

für $j = 2, ..., m$ wird die Monotoniebedingung erzwungen. Wie bei einer step-down Prozedur üblich, werden die Hypothesen $H_0^{r_1}, ..., H_0^{r_j}$ solange abgelehnt bis das erste Mal $\tilde{p}_{r_{j+1}} > \alpha$ eintritt. Die entsprechende sowie alle nachfolgenden Hypothesen $H_0^{r_{j+1}}, ...H_0^{r_m}$ können nicht mehr abgelehnt werden. (vgl. Zierer (2013): S.50-51)

step-down maxT Prozedur

Wie bei der single-step Variante gibt es auch hier das Analogon der maxT Prozedur (vgl. Dudoit u.a. (2003): S. 80):

$$\tilde{p}_{r_j} = \underset{t=1,...j}{max}\{\mathbb{P}(\underset{s\in\{r_t,...,r_m\}}{min} P_s \leq p_{r_t}|H_0^C)\} \ . \tag{21}$$

Die schrittweise Darstellung sieht hier für die b-te Resampling-Stichprobe, $b = 1, ..., B$, wie folgt aus: Im ersten Schritt werden die Teststatistiken $t_1^{*b}, ..., t_m^{*b}$ der b-ten Permutation für jede Hypothese H_0^j berechnet. Im zweiten Schritt werden dann die sukzessiven Maxima der Teststatistiken berechnet, $u_m^{*b} = |t_{r_m}^{*b}|$ und $u_j^{*b} = max(u_{j+1}^{*b}, |t_{r_j}^{*b}|)$ mit $j = 1, ..., m-1$, wobei r_j den Rang der beobachteten Teststatistiken bezeichnet, so dass $|t_{r_1}| \geq |t_{r_2}| \geq ... \geq |t_{r_m}|$ gilt. Die Monotonie der Resampling-Stichprobe ist dabei nicht zwingendermaßen dieselbe wie die der Originalstichprobe. Schließlich werden die adjustierten p-Werte durch

$$\tilde{q}_{r_j} = \frac{\sum_{b=1}^{B} \mathbb{P}(u_j^{*b} \geq |t_{r_j}|)}{B} \tag{22}$$

geschätzt. Mit Hilfe der sukzessiven Maxima

$$\tilde{p}_{r_1} = \tilde{q}_{r_1}, \ \tilde{p}_{r_j} = max(\tilde{q}_{r_j}, \tilde{p}_{r_{j-1}}) \tag{23}$$

für $j = 2, ..., m$ wird die Monotoniebedingung erzwungen. Bis zum ersten Mal $\tilde{p}_{r_{j+1}} > \alpha$ eintritt, werden alle Nullhypothesen $H_0^{r_1}, ..., H_0^{r_j}$ abgelehnt. Die zugehörige, sowie alle darauffolgenden Hypothesen $H_0^{r_{j+1}}, ..., H_0^{r_m}$ können nicht abgelehnt werden. (vgl. Zierer (2013): S. 51-52)

Beide minP Verfahren, sowie beide maxT Verfahren basieren auf der Annahme der globalen Nullhypothese H_0^C und stellen somit eine schwache Kontrolle der $FWER$ dar. Trifft jedoch die Subset Pivotality zu, so handelt es sich bei allen Prozeduren um starke Kontrollen der Fehlerrate. (vgl. Zierer (2013): S. 47-48 und S. 51) Diese Subset Pivotality ist folgendermaßen definiert: "Die Verteilung P hat die Subset Pivotality Eigenschaft, wenn, für alle Teilmengen $K \subseteq \{i; i \in J(\theta)\}$ von wahren Nullhypothesen, die gemeinsame Verteilung des Subvektors $\{P_i; i \in K\}$ unter $\bigcap_{i\in K} H_0^i$ und H_0^C identisch ist. Dabei bezeichnet $P = (P_1, ..., P_m)$ den Zufallsvektor der p-Werte" (Zierer (2013): S. 37).

Holm

Ein weiteres step-down Verfahren zur Kontrolle der $FWER$ ist von Holm (1979) und geht folgendermaßen vor: Finde ein $j^* = min\{j : p_{r_j} > \frac{\alpha}{m-j+1}\}$ und lehne alle Hypothesen H_{r_j} für $j = 1, ..., j^* - 1$ ab. Existiert so ein j^* nicht, lehne alle Hypothesen

ab. Die adjustierten p-Werte nach Holm sind definiert durch

$$\tilde{p}_{r_j} = \underset{t=1,...,j}{max}\{min((m-t+1)p_{r_t}, 1)\} \ . \tag{24}$$

Die Holm-Prozedur erzwingt durch das sukzessive Vorgehen eine Monotonie der adjustierten p-Werte $\tilde{p}_{r_1} \leq \tilde{p}_{r_2} \leq ... \leq \tilde{p}_{r_m}$. Somit kann eine einzelne Hypothese nur abgelehnt werden, wenn alle vorhergehenden Hypothesen, also alle mit kleineren nicht-adjustierten p-Werten, bereits abgelehnt wurden. (vgl. Dudoit u.a. (2003): S.79)

Hochberg

Das step-up Verfahren von Hochberg (1988) zur Kontrolle der $FWER$ ist das Pendant zum Verfahren von Holm. Es hat dieselben kritischen Werte, beginnt aber mit den größten p-Werten. Sei $j^* = max\{j : p_{r_j} \leq \frac{\alpha}{m-j+1}\}$, lehne alle Hypothesen H_{r_j} ab für $j = 1, ..., j^*$. Falls so ein j^* nicht existiert, lehne keine Hypothese ab. Die adjustierten p-Werte nach Hochberg sind somit definiert als

$$\tilde{p}_{r_j} = \underset{t=j,...,m}{min}\{min((m-t+1)p_{r_t}, 1)\} \ . \tag{25}$$

Da diese Prozedur ebenfalls sukzessiv vorgeht, erhält man auch hier eine Monotonie der adjustierten p-Werte. Vorteil des Verfahrens von Hochberg könnte sein, dass es mehr Power hat, da step-up Prozeduren oft mehr Power als ihren step-down Pendants zugeschrieben wird. Dazu aber später in Kapitel 4 mehr. (vgl. Dudoit u.a. (2003): S. 80)

Nachdem alle Adjustierungsverfahren zur Kontrolle der $FWER$ vorgestellt wurden, folgt nun die Beschreibung der Adjustierungsverfahren zur Kontrolle der FDR.

Benjamini&Hochberg

Das erste FDR-kontrollierende Verfahren unterliegt der Annahme, dass die Teststatistiken unabhängig sind. Diese step-up Prozedur von Bejamnini&Hochberg (1995) geht dabei folgendermaßen vor: Bestimme ein $j^* = max\{j : p_{r_j} \leq \frac{j}{m}\alpha\}$ und lehne alle Hypothesen H_{r_j} mit $j = 1, ..., j^*$ ab. Falls so ein j^* nicht existiert wird keine Hypothese abgelehnt. Die entsprechenden adjustierten p-Werte nach Benjamini&Hochberg sind, wie folgt, definiert:

$$\tilde{p}_{r_j} = \underset{t=j,...,m}{min}\{min(\frac{m}{t}p_{r_t}, 1)\} \ . \tag{26}$$

(vgl. Dudoit u.a. (2003): S. 80)

Benjamini&Yekutieli

Ein konservativeres Verfahren, das die FDR für willkürliche Abhängigkeitsstrukturen kontrolliert, kommt von Benjamini&Yekutieli (2001). Es handelt sich hierbei ebenfalls um eine step-up Prozedur, die mit folgender Definition für die adjustierten p-Werte

$$\tilde{p}_{r_j} = \min_{t=j,\ldots,m} \{min(\frac{m \sum_{j=1}^{m} \frac{1}{j}}{t} p_{r_t}, 1)\} \tag{27}$$

einen größeren Strafterm $\frac{m \sum_{j=1}^{m} \frac{1}{j}}{t}$ für große m hat als Benjamini&Hochberg mit $\frac{m}{t}$. (vgl. Dudoit u.a. (2003): S. 80-81)

4 Vergleich der Adjustierungsverfahren

In diesem Abschnitt sollen die soeben vorgestellten Methoden miteinander verglichen werden um dann im nächsten Kapitel, anhand des bereits vorgestellten Datenbeispiels, zu überprüfen, ob diese theoretischen Erkenntnisse auch zutreffen. Die einzigen zwei Ausnahmen bilden die single-step Varianten der minP und der maxT Prozeduren, die mangels bereits vorhandener Implementierung in R und ausreichender Zeit selbst eine zu machen, hier nicht zur Anwendung kommen werden.
Wie bereits erwähnt ist das Kriterium der $FWER$ strenger als das der FDR, somit werden Verfahren, die die $FWER$ kontrollieren, weniger Hypothesen ablehnen als Verfahren, die die FDR kontrollieren.
Vergleicht man nur $FWER$-kontrollierende Methoden wird Bonferroni als einzige single-step Prozedur das konservativste Verfahren sein. Denn hier gilt für alle p-Werte das gleiche strenge Kriterium $p_j < \frac{\alpha}{m}$. Im Vergleich dazu sind die Kriterien bei der step-down Prozedur nach Holm mit $p_{rj} < \frac{\alpha}{m-j+1}$ für kleine nicht-adjustierte p-Werte strenger als für große nicht-adjustierte p-Werte. Für $m = 5$ erhält man folgende Kriterien nach Bonferroni und Holm, die in Tabelle 3 aufgelistet sind.

j	Bonferroni	Holm
1	$\frac{\alpha}{5}$	$\frac{\alpha}{5}$
2	$\frac{\alpha}{5}$	$\frac{\alpha}{4}$
3	$\frac{\alpha}{5}$	$\frac{\alpha}{3}$
4	$\frac{\alpha}{5}$	$\frac{\alpha}{2}$
5	$\frac{\alpha}{5}$	$\frac{\alpha}{1}$

Tabelle 3: *Gleichbleibendes Kriterium beim konservativeren Verfahren von Bonferroni im Vergleich zum schwächer werdenden Kriterium bei Holm*

Folglich wird das Verfahren von Holm mehr Hypothesen ablehnen als das von Bonferroni. Hochberg hat dieselben kritischen Werte wie Holm, beginnt aber als step-up Prozedur mit den großen nicht-adjustierten p-Werten. Wie das folgende Beispiel in

Tabelle 4 zeigt kann es vorkommen, dass Hochberg eine Hypothese noch ablehnt, die Holm nicht mehr ablehnt:

	$p_{r_1}/\tilde{p}_{r_1}$	$p_{r_2}/\tilde{p}_{r_2}$	$p_{r_3}/\tilde{p}_{r_3}$	$p_{r_4}/\tilde{p}_{r_4}$	$p_{r_5}/\tilde{p}_{r_5}$
nicht-adjustiert	0.008	0.013	0.015	0.050	0.300
Holm	0.040	0.052	0.052	0.100	0.300
Hochberg	0.040	0.045	0.045	0.100	0.300

Tabelle 4: *Die Hypothesen 2 und 3 werden bei Hochberg noch abgelehnt bei Holm nicht mehr.*

Das ist auf die Monotonieeigenschaften der beiden Verfahren zurückzuführen. Denn bei Holm gilt $\tilde{p}_{r_2} = max(p_{r_1} \cdot m, p_{r_2} \cdot (m-1)) = max(0.040, 0.052) = 0.052$ und $\tilde{p}_{r_3} = max(p_{r_1} \cdot m, p_{r_2} \cdot (m-1), p_{r_3} \cdot (m-2)) = max(0.040, 0.052, 0.045) = 0.052$.
Und bei Hochberg gilt $\tilde{p}_{r_3} = min(p_{r_3} \cdot (m-2), p_{r_4} \cdot (m-3), p_{r_5} \cdot (m-4)) = min(0.045, 0.100, 0.300) = 0.045$ und $\tilde{p}_{r_2} = min(p_{r_2} \cdot (m-1), p_{r_3} \cdot (m-2), p_{r_4} \cdot (m-3), p_{r_5} \cdot (m-4)) = min(0.052, 0.045, 0.100, 0.300) = 0.045$. Ist die Anzahl an Nullhypothesen jedoch größer, ist der Unterschied nicht mehr so drastisch. (vgl. Zierer (2013): S. 54-55)
Schließlich bringt nach Dudoit das Zutreffen der Ungleichheit bei der step-down minP-Prozedur $\tilde{p}_{rj} = \max_{t=1,\ldots,j} \{P(\min_{s \in [rt,\ldots,rm]} P_s \leq p_{rt} | H_0^C)\}$ die p-Werte von Holm hervor und somit sind diese weniger konservativ als die Prozedur von Holm (vgl. Dudoit u.a. (2003): S. 80).
Der Unterschied zwischen den beiden FDR-kontrollierenden Verfahren basiert, wie bereits erwähnt, auf den unterschiedlichen Straftermen. Benjamini&Yekutieli haben mit $\frac{m \sum_{j=1}^{m} \frac{1}{j}}{t}$ den größeren Strafterm als Benjamini&Hochberg mit $\frac{m}{t}$ und sind somit konservativer.
Zusammengefasst wird erwartet, dass Bonferroni die wenigsten Hypothesen ablehnen wird, gefolgt von den schrittweisen Verfahren, die die $FWER$ kontrollieren. Dabei werden Holm und Hochberg auf sehr ähnliche Ergebnisse kommen, wobei bei Uneinigkeit Hochberg mehr signifikante Resultate ermitteln wird als Holm. Die beiden step-down Varianten von minP und maxT sind dabei weniger konservativ als Holm und Hochberg. Die meisten signifikanten adjustierten p-Werte sollten von den FDR-kontrollierenden Verfahren kommen. Allerdings wird die Prozedur von Benjamini&Yekutieli nicht ganz so viele Hypothesen ablehnen wie die von Benjamini&Hochberg.

5 Anwendung

In diesem Abschnitt werden nun die Methoden, die vorab behandelt wurden, angewendet und deren Ergebnisse miteinander verglichen. Als konkretes Datenbeispiel dienen die oben beschriebenen Daten. Da die Voraussetzungen für die Varianzanalyse scheinbar verletzt sind, werden die Ergebnisse des Kruskal-Wallis-Tests vorgezogen und näher

betrachtet. Die Ergebnisse des Shapiro-Wilk-Tests, sowie des Levene-Tests werden dann im Folgenden behandelt. Außerdem werden trotz der Annahmeverletzungen die Ergebnisse der Varianzanalyse kurz dargestellt und mit denen des Kruskal-Wallis-Tests verglichen.
Die Verfahren von Benjamini&Hochberg und Benjamini&Yekutieli werden in den folgenden Tabellen und Grafiken aus Platzgründen mit BH und BY abgekürzt.

5.1 Kruskal-Wallis-Test oder H-Test

Für den groben Überblick zeigt als erstes Tabelle 5 die Anzahlen der abgelehnten Hypothesen von insgesamt 200 Hypothesen jeder Adjustierungsmethode, sowie des nichtparametrischen Kruskal-Wallis-Tests, auch H-Test genannt, zu allen vier Messzeitpunkten. Wobei Hypothesen hier dann abgelehnt werden, wenn ihre p-Werte signifikant sind, was in diesen Fällen einen p-Werte kleiner gleich 0.05 bedeutet.

Zeitpunkt	Bonferroni	Holm	Hochberg	minP	maxT	BY	BH	H-Test
0	148	160	160	163	163	172	184	184
1	141	159	159	163	163	167	179	180
2	141	150	150	154	155	165	179	179
3	136	146	146	153	153	164	178	178

Tabelle 5: *Die Anzahl signifikanter p-Werte von insgesamt 200 Hypothesen des Kruskal-Wallis-Tests bzw. H-Tests und der verschiedenen Adjustierungsmethoden zu den Zeitpunkten 0, 1, 2 und 3*

Wie im Vergleichsteil vorausgesagt, werden dem Verfahren von Bonferroni zu Folge die wenigsten Hypothesen abgelehnt, gefolgt von den schrittweisen Verfahren von Holm und Hochberg, die sich hier durchweg einig sind. Auch die beiden Resampling-Verfahren minP und maxT erzielen sehr ähnliche Ergebnisse. Insgesamt sind die Prozeduren zur Kontrolle der FDR weniger konservativ als die $FWER$- kontrollierenden Verfahren. Dabei lehnt die Methode von Benjamini&Yekutieli immer weniger Hypothesen ab als jene von Benjamini&Hochberg. Das Verfahren von Benjamini&Hochberg ermittelt sogar fast genauso viele bis exakt genauso viele signifikante Hypothesen wie der Kruskal-Wallis-Test ohne Adjustierungen. Über die Zeit hinweg betrachtet kommen die Verfahren auf sehr ähnliche Resultate. Wobei alle tendenziell weniger signifikante Ergebnisse ermitteln, je später die Messung durchgeführt wurde.

Um den Verlauf der ansteigenden p-Werte erkennen zu können, folgen ein paar Streudiagramme in den Abbildungen 1-4. Diese stellen die nicht-adjustierten p-Werte des Kruskal-Wallis-Tests auf der x-Achse und die adjustierten p-Werte auf der y-Achse dar. So kann verfolgt werden wie sich die adjustierten p-Werte mit den ansteigenden nicht-adjustierten p-Werten verhalten. Bei 0.05 sind jeweils rot gestrichelte Linien eingezeichnet, um anzuzeigen wann eine Hypothese noch als signifikant angesehen wird

und ab wann nicht mehr. Zur besseren Darstellung der durch die Adjustierung nicht mehr signifikanten p-Werte, die ohne Adjustierung noch signifikant wären, sind in weiteren Streudiagrammen auf der x-Achse nur noch die p-Werte des Kruskal-Wallis-Tests bis 0.05 angezeigt.

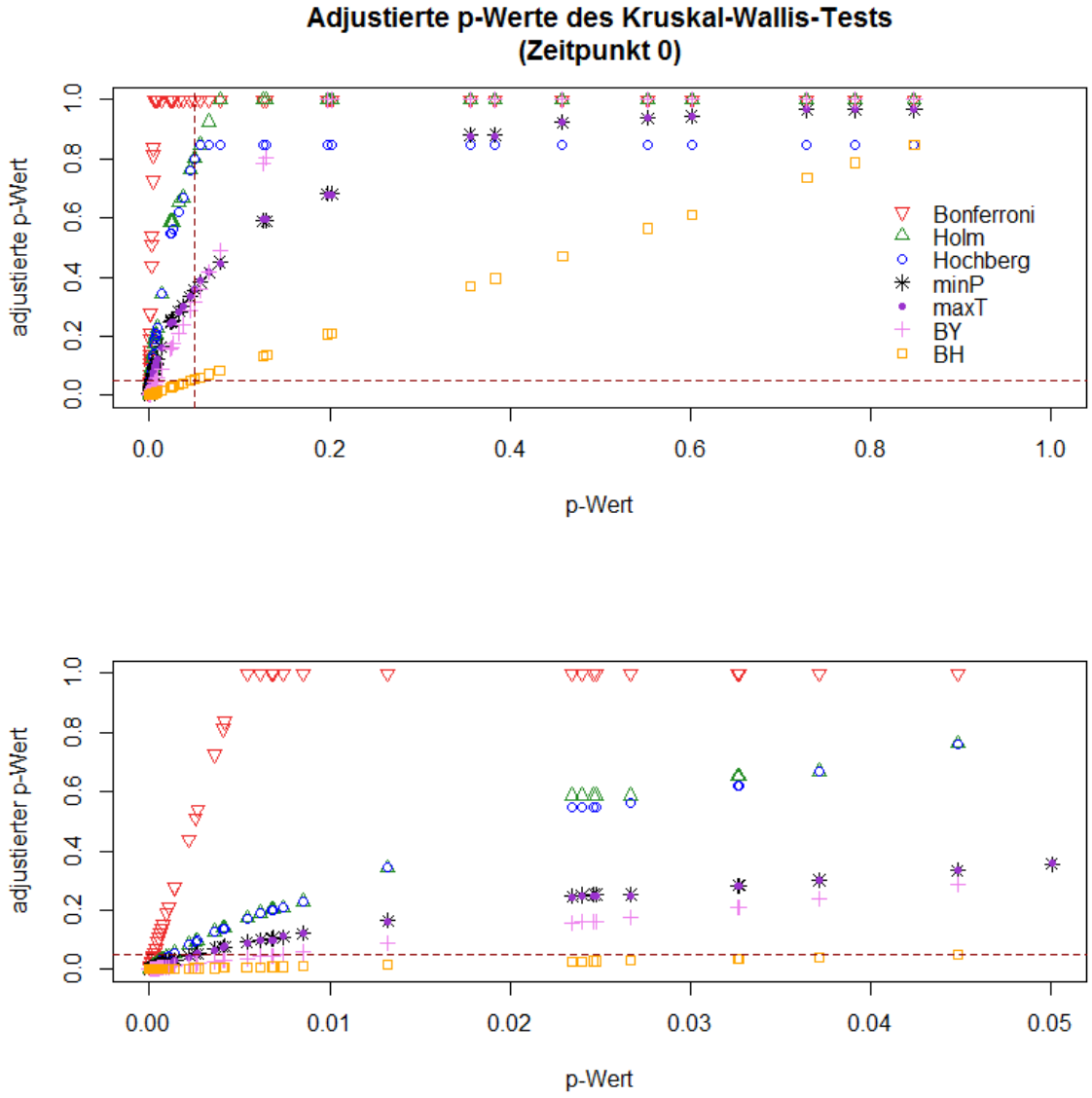

Abb. 1: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte des Kruskal-Wallis-Tests abgebildet und unten nur die p-Werte des Kruskal-Wallis-Tests bis 0.05.*

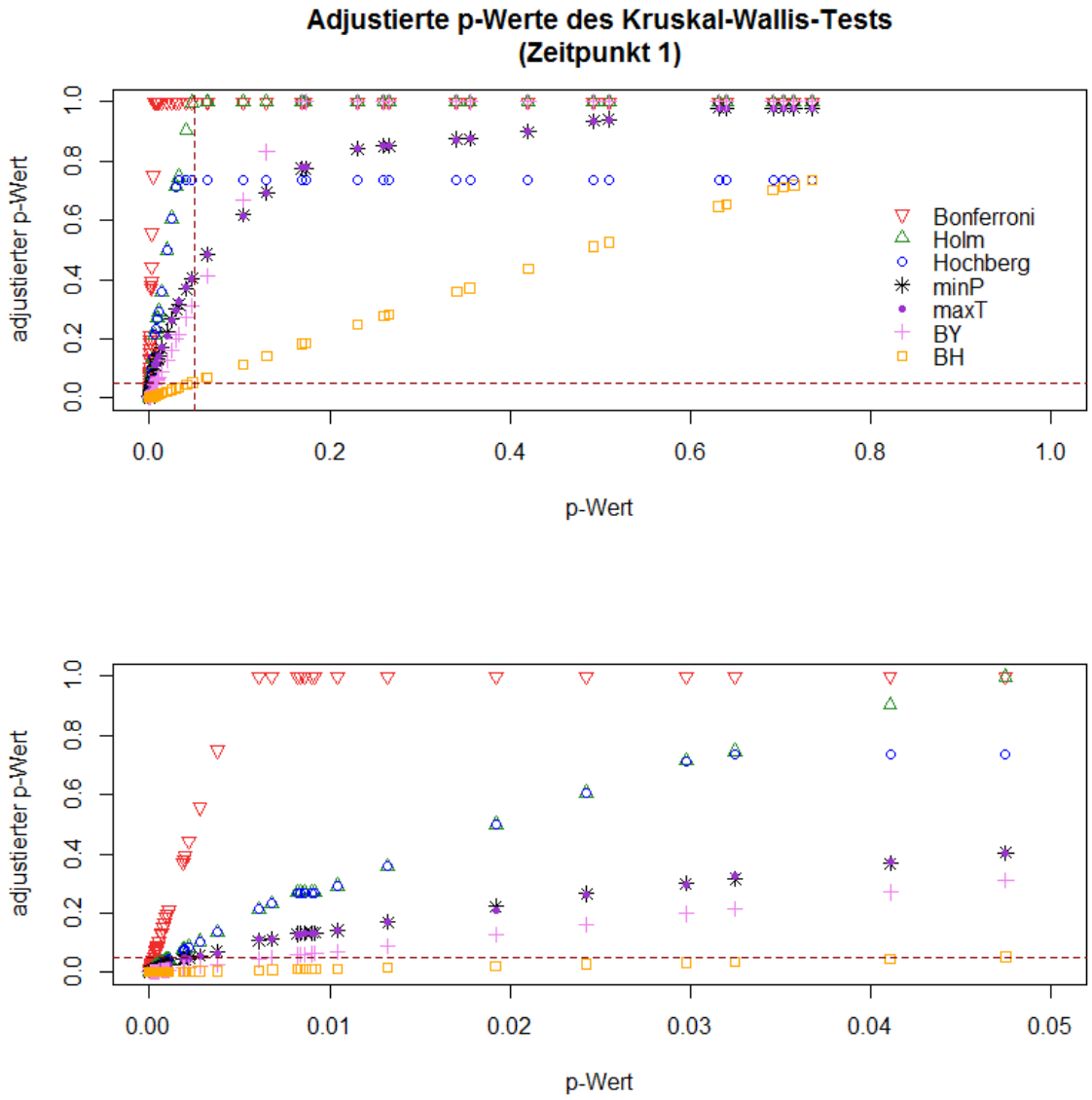

Abb. 2: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte des Kruskal-Wallis-Tests abgebildet und unten nur die p-Werte des Kruskal-Wallis-Tests bis 0.05.*

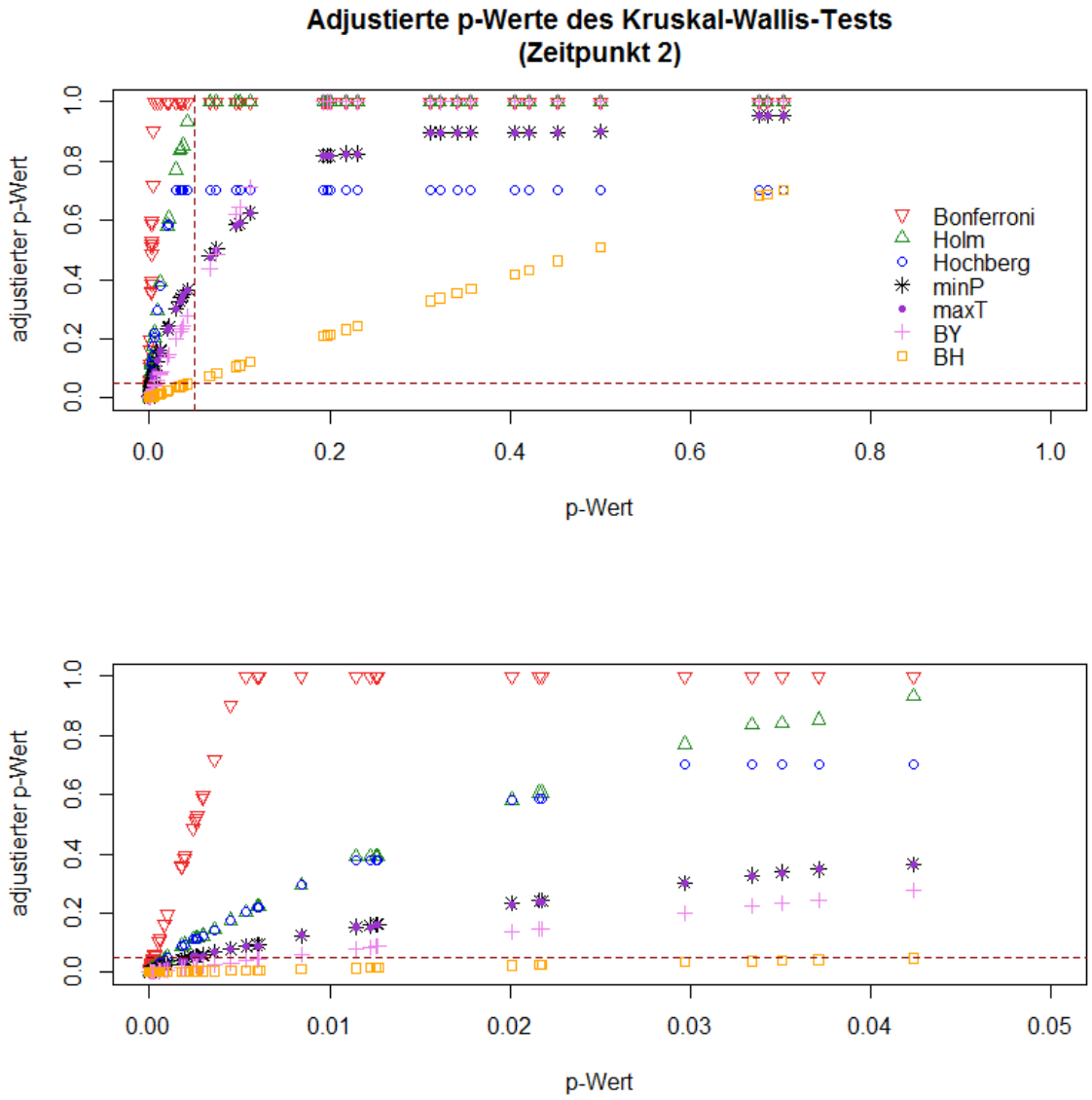

Abb. 3: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte des Kruskal-Wallis-Tests abgebildet und unten nur die p-Werte des Kruskal-Wallis-Tests bis 0.05.*

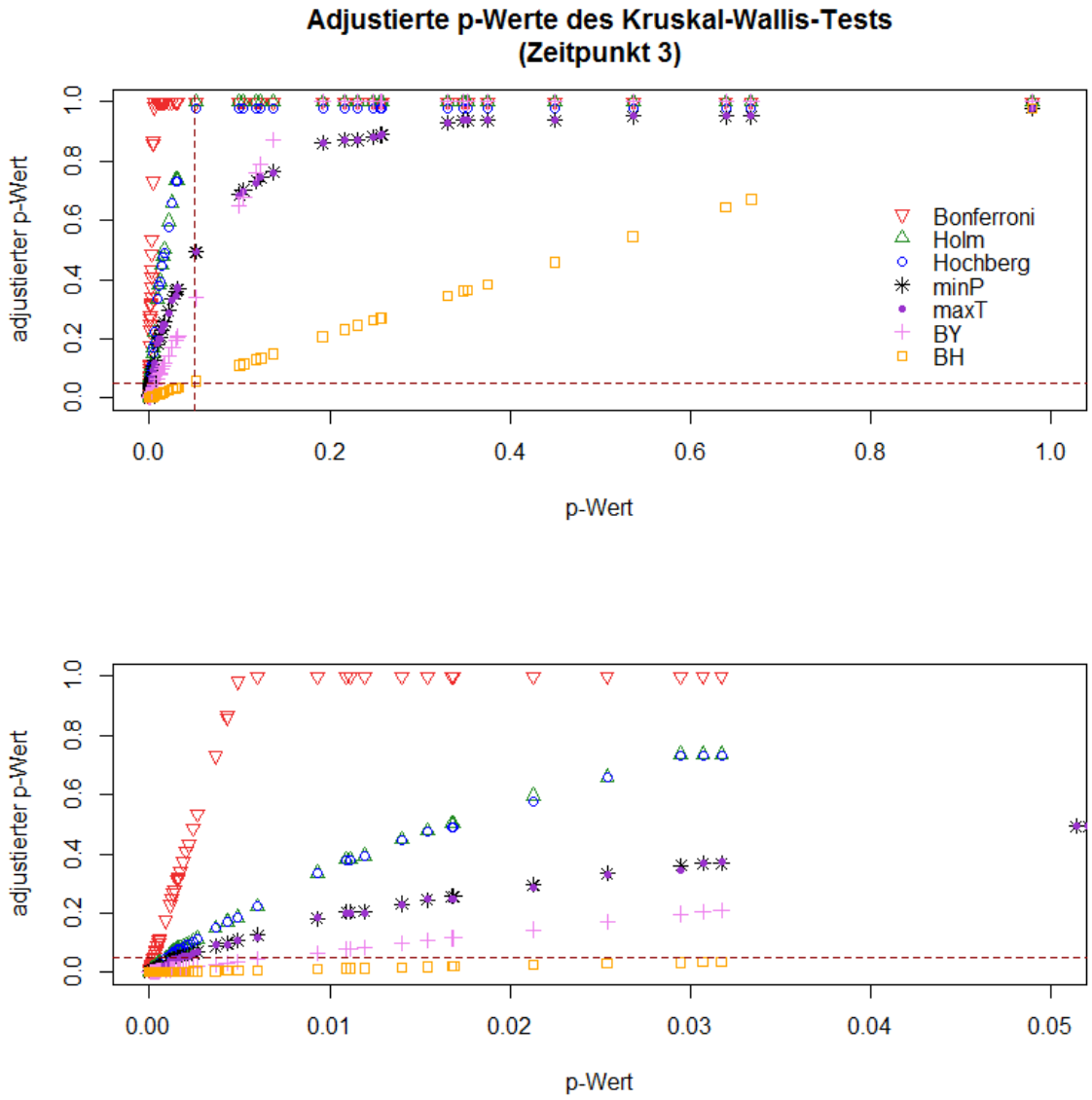

Abb. 4: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte des Kruskal-Wallis-Tests abgebildet und unten nur die p-Werte des Kruskal-Wallis-Tests bis 0.05.*

Die Kurven von Bonferroni, Holm und Hochberg haben alle einen ähnlichen Verlauf mit einem deutlichen Knick. Bonferroni läuft dabei am steilsten auf die 1.0 zu. Holm und Hochberg liegen lange auf derselben Kurve, wobei Hochberg früher abknickt und nicht bis zur 1.0 ansteigt und Holm letztlich auch bei der 1.0 landet. Da die beiden sich erst bei p-Werten des Kruskal-Wallis-Tests um die 0.05 herum trennen, wo beide sich schon längst außerhalb der signifikanten Werte befinden, macht sich dieser Unterschied in der Anzahl der abgelehnten Hypothesen nicht bemerkbar. Die Resampling-Methoden haben ebenfalls einen gemeinsamen Verlauf und trennen sich nie deutlich voneinander. Ihre Kurve beginnt flacher als die vorhergehenden und hat keinen so extremen Knick. Die erste FDR-kontrollierende Prozedur von Benjamini&Yekutieli hat einen ähnlichen Verlauf wie Bonferroni und Holm, beginnt nur wesentlich flacher, bleibt somit länger unter 0.05 und landet deutlich später bei der 1.0. Die adjustierten p-Werte von Benjamini&Hochberg folgen eher einer gleichmäßig ansteigenden Geraden, die am flachsten von allen beginnt. Folglich ergeben sich hier die meisten adjustierten p-Werte, die kleiner gleich 0.05 sind.

Um auch der inhaltlichen Frage nachzukommen anhand welcher Moleküle sich die drei Erregertypen voneinander unterscheiden lassen, seien folgende Tabellen dargestellt. Für jedes Molekül und jedes Verfahren wird angezeigt, ob ein signifikantes Ergebnis beobachtet wurde oder nicht. 0 steht hier für p-Werte kleiner gleich 0.05, also signifikante Resultate und 1 für p-Werte größer 0.05, also nicht signifikant. In jeder Spalte befinden sich vier Zahlen, die erste steht für den Messzeitpunkt 0, die zweite für den Messzeitpunkt 1 und so weiter.
In Tabelle 6 sind nur die Moleküle enthalten, die bei allen Verfahren zu jedem Zeitpunkt einen signifikanten p-Wert haben. Dies trifft auf insgesamt 125, also mehr als die Hälfte, der Moleküle zu. In diesen Fällen würde es sich wohl lohnen der Frage, welche Erregergruppen sich genau voneinander unterscheiden, weiter nachzugehen. Nachdem diese Tabelle nur Nuller enthalten würde und mit 125 Zeilen sehr lang wäre, sind nur die Namen der Moleküle in Tabelle 6 aufgelistet.

Namen der Moleküle mit ”0 0 0 0” bei allen Verfahren

CH4.	X17.	X20.	X21.	Acetylene.	Methanol.	O2.33..
X35.	X38.	X39.	ACN.	X42.	X43.	X45.
Formic.Acid.	X50.	X51.	X52.	X53.	X54.	X55.
X56.	X57.	X58.	X61.	X62.	X63.	SO2.
X65.	X66.	X67.	X72.	X73.	X76.	X77.
Benzene.Xe.	X80.	X81.	X86.	X87.	X88.	X89.
X90.	X91.	X92.	X93.	X94.	X100.	X101.
X102.	X103.	X104.	X107.	X108.	X109.	X110.
X112.	X113.	X114.	X115.	X116.	X117.	X118.
X119.	X120.	X121.	X122.	EI.H2	NH3	M19

Namen der Moleküle mit "0 0 0 0" bei allen Verfahren						
Ethylene	M29	NO	CH3NH2	M33	H2S	M36
M37	M38	M40	M41	M43	Acetaldehyde	Ethanol
M48	M50	M51	Butadiene	M55	M56	M57
M61	M62	M63	M64	M67	Isoprene	M69
M73	M74	M75	M76	M80	M81	M87
M90	M91	Toluene	M93	M94	M98	M101
M103	M108	M109	M115	M116	M117	M118
M119	M120	M121	M122	M123	M135	

Tabelle 6: *Auflistung aller Moleküle, die über alle Zeitpunkte, beim Kruskal-Wallis-Test, sowie allen Adjustierungsmethoden immer signifikant sind.*

Umgekehrt sind in Tabelle 7 nur Moleküle dargestellt, bei denen sich nie ein signifikantes Resultat ergibt. Insgesamt handelt es sich dabei um neun Moleküle, für die sich eine weiterführende Analyse wohl kaum lohnen wird, da man doch recht sicher davon ausgehen kann, dass die Messwerte dieser Moleküle sich nicht signifikant zwischen den Gruppen der gram negativen, der gram positiven und der Pilze unterscheiden.

Molekül	**Bonferroni**	**Holm**	**Hochberg**	**minP**	**maxT**	**BY**	**BH**	***H*-Test**
X59.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M49	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
Propanol	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M60	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M82	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M85	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M88	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M96	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1
M111	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1

Tabelle 7: *Zusammenfassung aller Ergebnisse des Kruskal-Wallis-Tests bzw. H-Tests, die über alle Zeitpunkte und bei allen Ajustierungsmethoden nicht signifikant sind; 0 steht für signifikant und 1 steht für nicht signifikant.*

Und bei den letzten 66 Fällen, die in Tabelle 8 dargestellt sind, finden sich je nach Messzeitpunkt und angewandtem Verfahren mal signifikante, mal nicht signifikante Ergebnisse. Hier sind die Resultate also nicht so eindeutig, dass sich alle Verfahren einig sind. Außerdem wird hier deutlich, weshalb die Rede von volatilen, also flüchtigen, Komponenten ist. Da sich die Messwerte von manchen Molekülen nur zu bestimmten Zeitpunkten wesentlich voneinander unterscheiden, scheinen diese organischen Komponenten nicht in stabiler Form vorhanden zu sein, sondern teilweise erst zu entstehen oder sich wieder abzubauen.

Molekül	Bonferroni	Holm	Hochberg	minP	maxT	BY	BH	H-Test
M27.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
M29..	0 1 1 1	0 1 1 1	0 1 1 1	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0
Formaldehyde.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 0 0	0 0 0 0
X40.	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
N2O.	0 1 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
HNO2.	0 1 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
X48.	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X49.	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
Acetic.Acid.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0
X68.	0 1 1 1	0 1 1 1	0 1 1 1	0 1 1 1	0 1 1 1	0 1 1 1	0 0 0 1	0 0 0 1
X69.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
X70.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1	0 1 1 1
X71.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	0 1 1 0	0 0 1 0
X74.	1 1 1 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X75.	0 1 1 1	0 0 1 1	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X79.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 0 0	0 0 0 0
X82.	1 1 1 1	0 1 1 1	0 1 1 1	0 1 1 1	0 1 1 1	0 0 0 0	0 0 0 0	0 0 0 0
X83.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 0	0 0 0 0	0 0 0 0
X84.	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X85.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 0 0	0 0 0 0
X95.	0 1 1 1	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 0 0 0	0 0 0 0
X96.	0 1 1 1	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0
X97.	0 1 1 1	0 1 1 1	0 1 1 1	0 1 1 0	0 1 1 0	0 0 0 0	0 0 0 0	0 0 0 0
X98.	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X99.	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X105.	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
X106.	0 1 1 1	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 0 0 0	0 0 0 0	0 0 0 0
X111.	0 1 1 1	0 0 1 1	0 0 1 1	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0
EI.H2....M1	1 1 1 1	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
EI.H2O.18	1 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
EI.N2.28	1 1 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0
EI.O2.32	1 1 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
EI.CO2.44	1 0 0 1	1 0 0 1	1 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M35	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 1 1	0 0 1 1
M39	1 1 1 0	1 1 1 0	1 1 1 0	1 0 1 0	1 0 1 0	1 0 0 0	1 0 0 0	1 0 0 0
Propene	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	0 0 0 0	0 0 0 0
M46	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 0 0 0	0 0 0 0	0 0 0 0
M47	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 0 1	1 1 0 1
M52	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M53	0 1 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0
Acetone	1 1 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0

Molekül	Bonferroni	Holm	Hochberg	minP	maxT	BY	BH	H-Test
M65	1 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
M66	0 1 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0
M70	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0
M71	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 0 1	0 0 0 0	0 0 0 0
M72	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0
M77	1 1 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
Benzene	1 1 0 1	1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 0	1 0 0 0
M79	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 0 1	0 1 0 1	0 0 0 1	0 0 0 1
M83	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
M84	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 1 0	0 0 1 0
M86	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 0 1	0 0 0 0	0 0 0 0
M89	1 1 1 1	1 1 1 1	1 1 1 1	1 0 1 1	1 0 1 1	1 0 1 1	1 0 0 1	1 0 0 1
M95	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 0 0	1 1 0 0
M97	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0
M99	1 1 1 1	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0
M100	0 1 1 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M102	1 1 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M104	0 0 1 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0
M105	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1
M106	1 1 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1
M107	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 0	0 1 1 0	0 1 1 0
M110	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
M112	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
M113	1 1 1 1	1 1 1 1	1 1 1 1	1 0 1 1	1 0 1 1	1 0 1 1	0 0 0 0	0 0 0 0
M114	1 0 1 1	1 0 0 1	1 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0

Tabelle 8: *Zusammenfassung aller Ergebnisse des Kruskal-Wallis-Tests bzw. H-Tests, die je nach Messzeitpunkte und Adjustierungsmethode signifikant oder nicht signifikant sind; 0 steht für signifikant und 1 steht für nicht signifikant.*

5.2 Varianzanalyse

Obwohl beim Shapiro-Wilk-Test und beim Levene-Test herauskam, dass beim Großteil der Moleküle die Normalverteilungsannahme bzw. die Annahme gleicher Varianzen zwischen den Erregergruppen offenbar verletzt ist, sollen hier die Ergebnisse der Varianzanalyse kurz zusammengefasst werden und anschließend mit den Ergebnissen des Kruskal-Wallis-Tests verglichen werden.

Zunächst werden die Ergebnisse zur Überprüfung der Testannahmen betrachtet. Zur Veranschaulichung sind die p-Werte der beiden Tests zum Messzeitpunkt 0 in Abbildung 5 grafisch dargestellt. Beim Shapiro-Wilk-Test sind insgesamt 14 p-Werte größer 0.05, das heißt nur die Messwerte von 14 Molekülen sind zum Zeitpunkt 0 nor-

malverteilt. Zu den anderen Zeitpunkten ergibt sich ein ähnliches Bild. Zum Zeitpunkt 1 sind es 30 nicht signifikante Ergebnisse beim Shapiro-Wilk-Test. Und zu den Zeitpunkten 2 und 3 liegen 24 und 23 normalverteilte Moleküle vor. Von insgesamt 200 Molekülen ist also nur ein geringer Anteil normalverteilt.

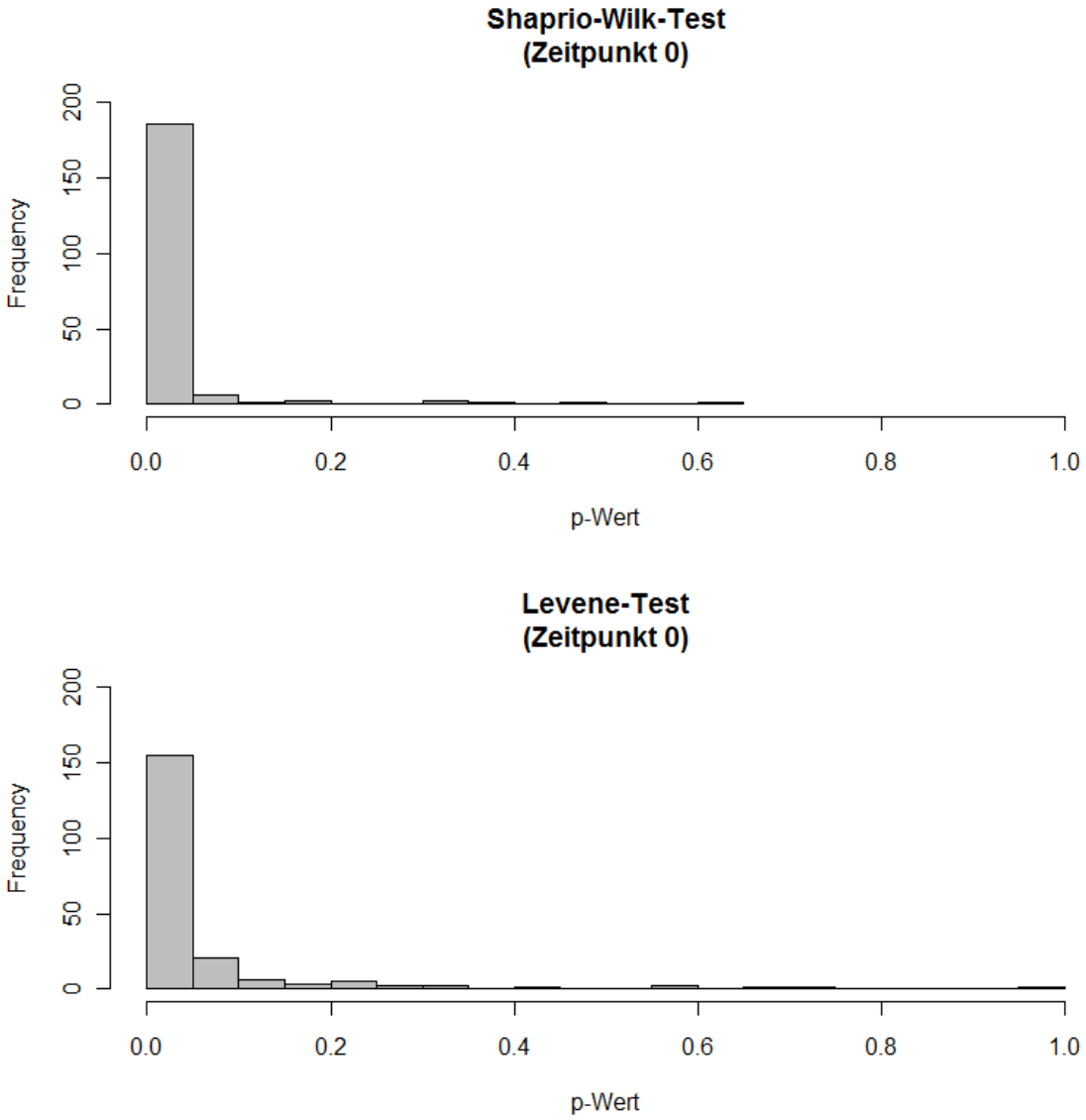

Abb. 5: *Histogramme der p-Werte des Shapiro-Wilk-Tests (oben) und des Levene-Tests (unten), jeweils der erste Balken stellt die Moleküle dar, die die jeweilige Annahme verletzen.*

Auch die Homoskedastizität ist überwiegend verletzt. Zum Messzeitpunkt 0 liegen lediglich 45 p-Werte des Levene-Tests über 0.05. Es finden sich 63 homoskedastische Moleküle zum Zeitpunkt 1. Und schließlich liegen zu den Zeitpunkten 2 und 3 37 und 51 Moleküle mit Varianzhomogenität zwischen den Erregertypen vor.

Zusammenfassend ist sowohl die Annahme der Normalverteilung als auch die der Homoskedastizität verletzt, so dass die Ergebnisse der Varianzanalyse mit Vorsicht zu genießen sind. Trotzdem ist diese hier gerechnet worden und ihre Ergebnisse wer-

den anschließend gezeigt. Zuvor wird aber noch dargestellt, wie die Verteilungen der Moleküle stattdessen aussehen. Da hier nicht 200 Histogramme, Boxplots, QQ-Plots oder andere Grafiken gezeigt werden können um die Verteilungen zu beschreiben, wird versucht mit Hilfe von verschiedenen Maßzahlen ein paar Eigenschaften der Verteilungen der Moleküle zusammenzufassen. Die Darstellung der Maßzahlen bezieht sich immer auf ein beliebiges Molekül j, wobei darauf verzichtet wird überall den Index j hinzuschreiben.
Ein Merkmal von Verteilungen ist die Symmetrie bzw. Schiefe. Man unterscheidet zwischen symmetrischen, linkssteilen und rechtssteilen Verteilungen. Bei symmetrischen Verteilungen sind die linke und die rechte Hälfte der Verteilung annähernd spiegelbildlich. Linkssteile Verteilungen haben den überwiegenden Teil ihrer Daten linksseitig und bei rechssteilen Verteilungen ist der Großteil der Daten entsprechend rechtsseitig. Eine Möglichkeit die Schiefe zu beurteilen bietet der Quantilskoeffizient der Schiefe, der im Gegensatz zum Momentkoeffizient resistent gegen Ausreißer ist und deshalb hier verwendet wird. Der Quantilskoeffizient hat folgende Formel:

$$g_p = \frac{(x_{1-p} - \tilde{x}) - (\tilde{x} - x_p)}{x_{1-p} - x_p} \ , \tag{28}$$

die im Zähler den Unterschied zwischen der Entfernung des p-Quantils und der des $(1-p)$-Quantils jeweils zum Median $\tilde{x}$ misst. Da bei linkssteilen Verteilungen das untere Quantil näher am Median ist und bei rechtssteilen weiter entfernt liegt vom Median, gilt:
$g_p = 0$ für symmetrische Verteilungen,
$g_p > 0$ für linkssteile Verteilungen und
$g_p < 0$ für rechtssteile Verteilungen.
Ist $p = 0.25$ erhält man den Quartilskoeffizienten. (vgl. Farhmeir u.a. (2010): S. 48 und S.74-75)
Wie die Histogramme der Quartilskoeffizienten der vier Messzeitpunkte in Abbildung 6 (oben) zeigen, sind symmetrische, linkssteile und rechtssteile Verteilungen alle vertreten. Es sind jedoch mehr linkssteile Verteilungen als rechtssteile und unter den linkssteilen sind auch extremer ausgeprägte Verteilungen.

Ein weiteres Merkmal von Verteilungen ist die Wölbung, auch Kurtosis genannt. Diese gibt an wie stark der zentrale Bereich bzw. die Enden der Daten besetzt sind. Ist eine Verteilung in der Mitte eher spitz, sind die Enden stärker besetzt als bei Verteilungen, die in der Mitte flacher sind. Als Vergleich dafür, was breit oder spitz bedeutet, dient die Normalverteilung. Das Wölbungsmaß von Fisher, das als Maßzahl für die Kurtosis dient, ist so definiert, dass es bei Normalverteilungen gleich Null ist:

$$\gamma = \frac{m_4}{s^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3 \ . \tag{29}$$

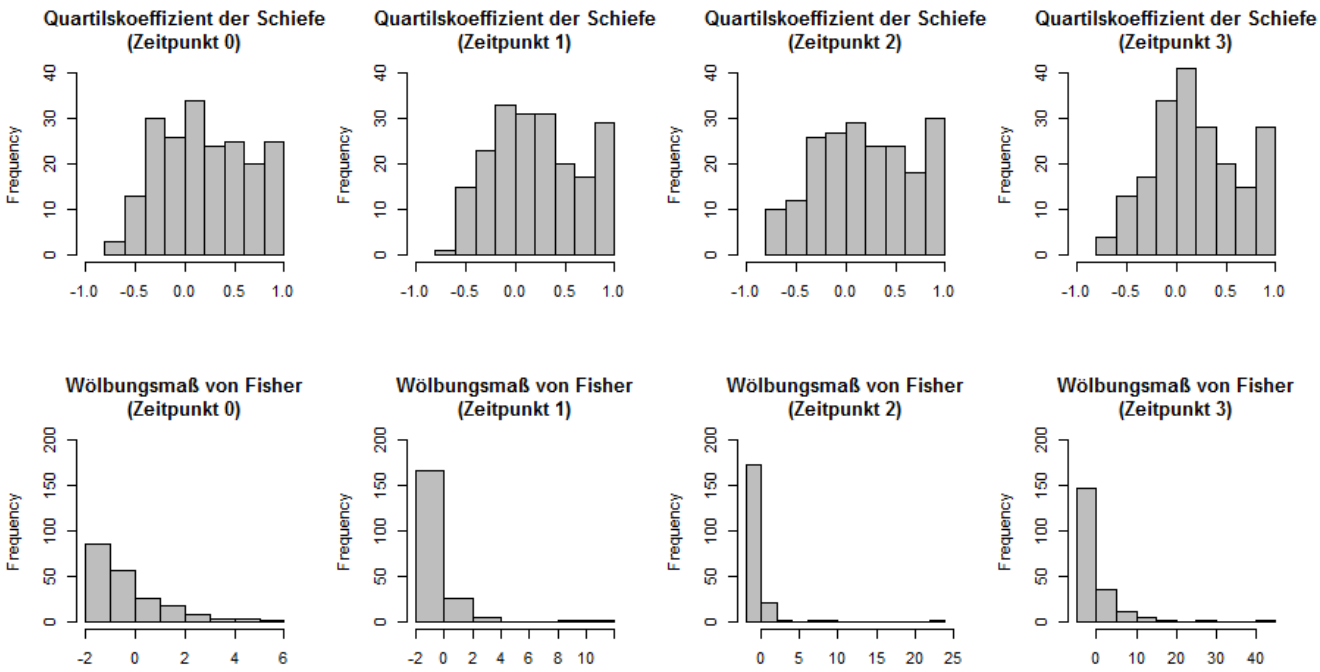

Abb. 6: *Oben sind die Werte des Quartilskoeffizienten der Schiefe dargestellt und unten die Werte des Wölbungsmaßes von Fisher.*

Dabei ist s^4 die quadrierte Stichprobenvarianz und es gilt:
$\gamma = 0$ bei Normalverteilung,
$\gamma > 0$ bei spitzeren Verteilungen und
$\gamma < 0$ bei flacheren Verteilungen. (vgl. Fahrmeir u.a. (2010): S. 76)
Bei Betrachtung der Histogramme des Wölbungsmaßes in Abbildung 6 (unten) wird deutlich, dass der Großteil der Moleküle flachere Verteilungen hat als die Normalverteilung. Wobei beachtet werden muss, dass die vier Grafiken zu den vier Messzeitpunkten verschieden skalierte x-Achsen haben, also optisch nicht direkt vergleichbar sind. Es ist trotzdem eindeutig erkennbar, dass alle Histogramme eine linkssteile Verteilung zeigen, also mehr kleine Werte von γ zu beobachten sind.

Und schließlich können mögliche Ausreißer eine Verteilung charakterisieren. Um potenzielle Ausreißer, also Datenpunkte, die weit entfernt von den anderen Daten liegen, zu ermitteln verwendet man häufig den Interquartilsabstand $d_Q = x_{0.75} - x_{0.25}$. Wobei $x_{0.75}$ das 75%-Quantil und $x_{0.25}$ das 25%-Quantil bezeichnet. Dieser Interquartilsabstand dient als Maßzahl für die Streuung von Daten. Liegen Datenpunkte außerhalb eines sogenannten Zauns, der anhand des Interquartilsabstands berechnet wird, gelten sie, einer Faustregel nach, als potenzielle Ausreißer. Dazu gehören also Punkte, die kleiner als die Untergrenze $z_u = x_{0.25} - 1.5d_Q$ oder größer als die Obergrenze $z_o = x_{0.75} + 1.5d_Q$ sind. (vgl. Fahrmeir u.a. (2010): S. 66-67)
Tabelle 9 zeigt wie viele potenzielle Ausreißer oder zumindest Extrempunkte nach dieser Faustregel zu den vier Messzeitpunkten bei den 200 Molekülen gefunden wur-

den. Dabei gibt es insgesamt 67 Beobachtungen je Molekül zum Messzeitpunkt 0 und 68 zu den anderen drei Messzeitpunkten.

Zeitpunkt\Ausreißer	0	1	2	3	4	5	6	7	8	9	10	11	12
0	144	11	4	1	24	3	7	1	0	0	2	0	3
1	163	11	7	4	6	0	5	4	0	0	0	0	0
2	167	11	6	2	4	2	3	0	3	0	1	1	0
3	142	11	11	7	12	7	3	1	3	1	0	1	1

Tabelle 9: *Anzahl potenzieller Ausreißer je Messzeitpunkt von 67 Beobachtungen je Molekül zum Messzeitpunkt 0 und 68 Beobachtungen je Molekül zu den restlichen Messzeitpunkten, bei insgesamt 200 Molekülen*

Es sind auffälligere Messwerte bei einigen Molekülen zu beobachten, der Großteil hat jedoch keine oder nur wenige potenzielle Ausreißer.

Nun kommen die Ergebnisse der Varianzanalyse und der anschließenden Adjustierungen, die in Tabelle 10 zusammengefasst sind. Von insgesamt 200 Molekülen haben die Verfahren folgende Anzahlen an abgelehnten Hypothesen ergeben.

Zeitpunkt	Bonferroni	Holm	Hochberg	minP	maxT	BY	BH	Anova
0	151	156	156	161	161	165	178	178
1	153	165	165	166	166	168	178	180
2	149	156	156	161	162	168	174	175
3	141	146	146	153	153	162	177	178

Tabelle 10: *Anzahl signifikanter p-Werte der 200 Hypothesen bei der Varianzanalyse bzw. Anova und der verschiedenen Adjustierungsmethoden zu den Zeitpunkten 0, 1, 2 und 3*

Die Ergebnisse sind sehr ähnlich zu denen des Kruskal-Wallis-Tests. Die Varianzanalyse lehnt etwas weniger Hypothesen ab, genauso das Verfahren von Benjamini&Hochberg. Einzig das Verfahren von Bonferroni lehnt hier durchweg mehr Hypothesen ab als beim Kruskal-Wallis-Test. Die restlichen Verfahren haben keine so eindeutige Tendenz. Der zeitliche Aspekt, dass je später die Messung durchgeführt wird, desto weniger signifikante Hypothesen beobachtet werden, ist hier nicht ganz so klar zu erkennen.
Auch hier werden die adjustierten p-Werte der verschiedenen Methoden in Abhängigkeit der p-Werte der Varianzanalyse in den Abbildungen 7-10 grafisch dargestellt. Wie nach Betrachtung der Anzahlen der abgelehnten Hypothesen zu erwarten, ergibt sich ein vergleichbares Bild wie bei den p-Werten des Kruskal-Wallis-Tests und deren adjustierten p-Werte.

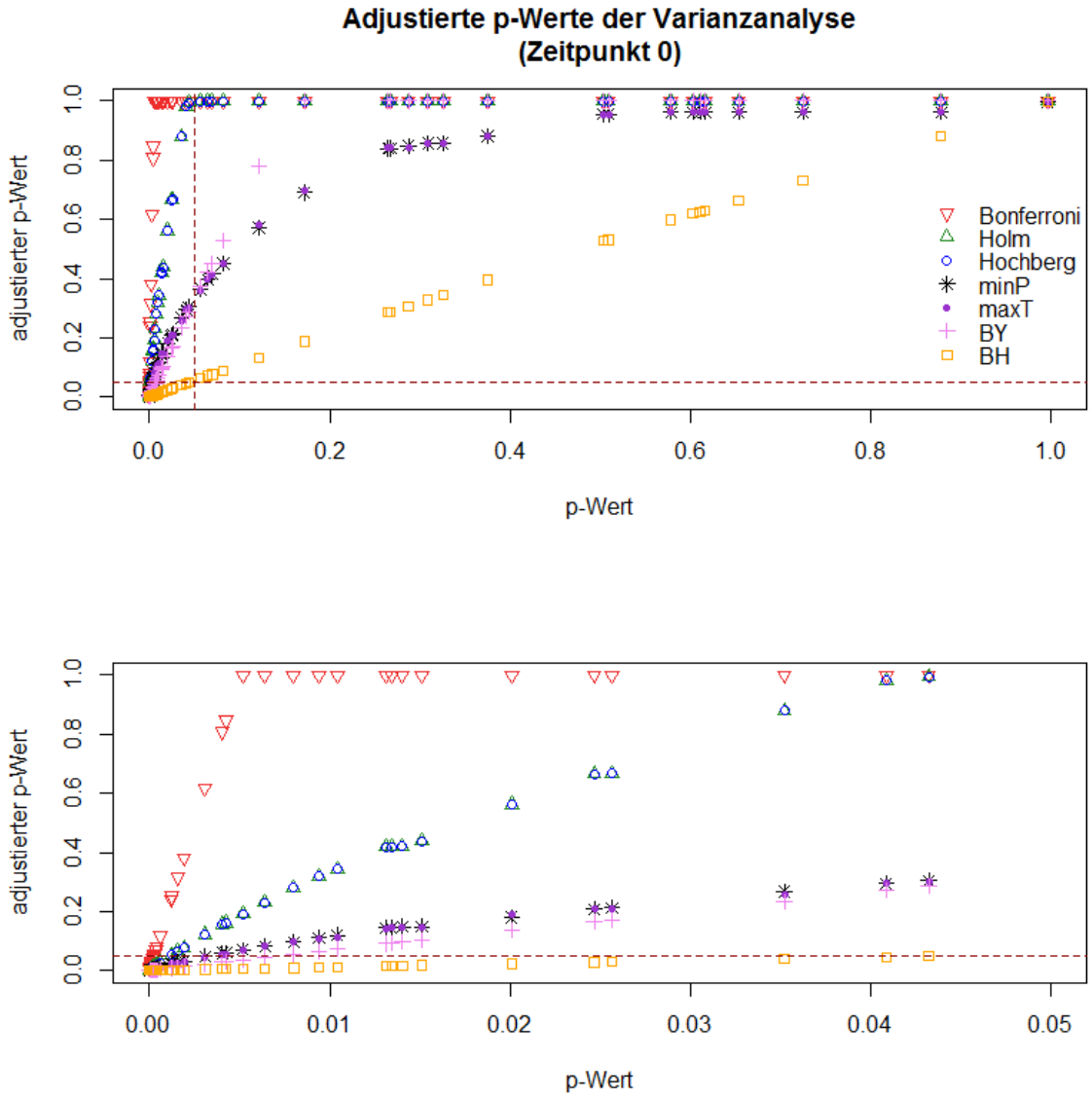

Abb. 7: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte der Varianzanalyse abgebildet und unten nur die p-Werte der Varianzanalyse bis 0.05.*

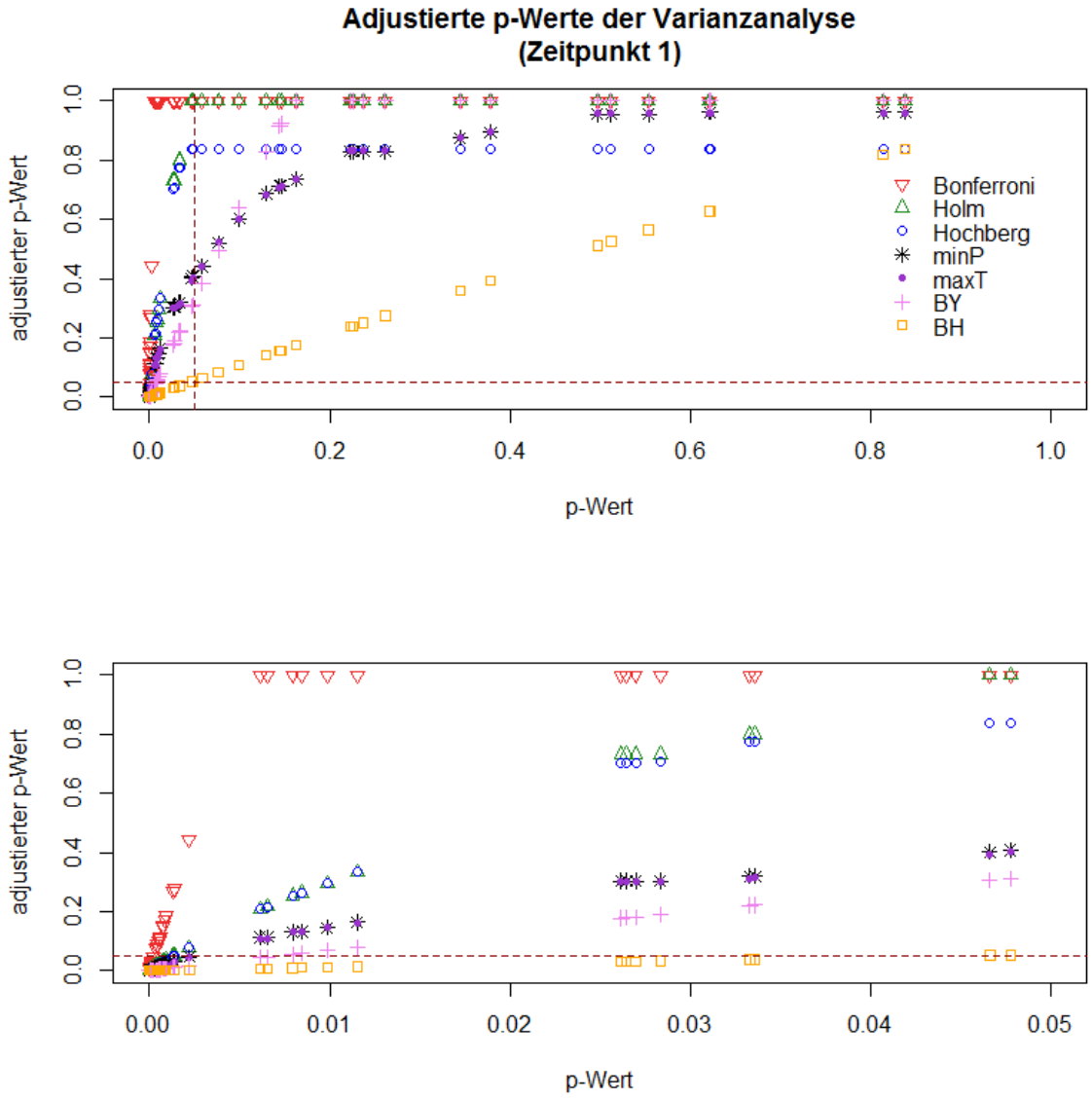

Abb. 8: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte der Varianzanalyse abgebildet und unten nur die p-Werte der Varianzanalyse bis 0.05.*

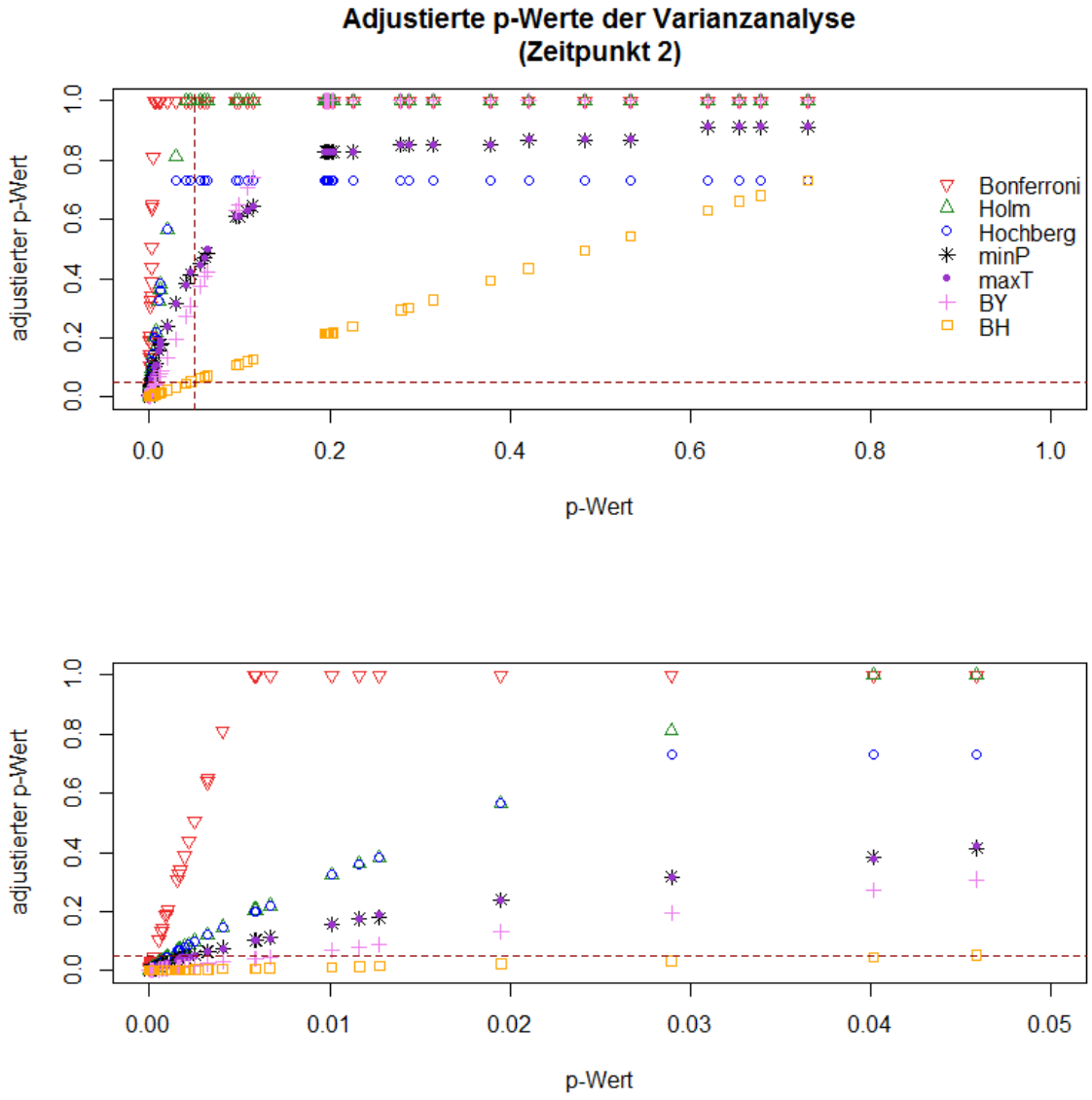

Abb. 9: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte der Varianzanalyse abgebildet und unten nur die p-Werte der Varianzanalyse bis 0.05.*

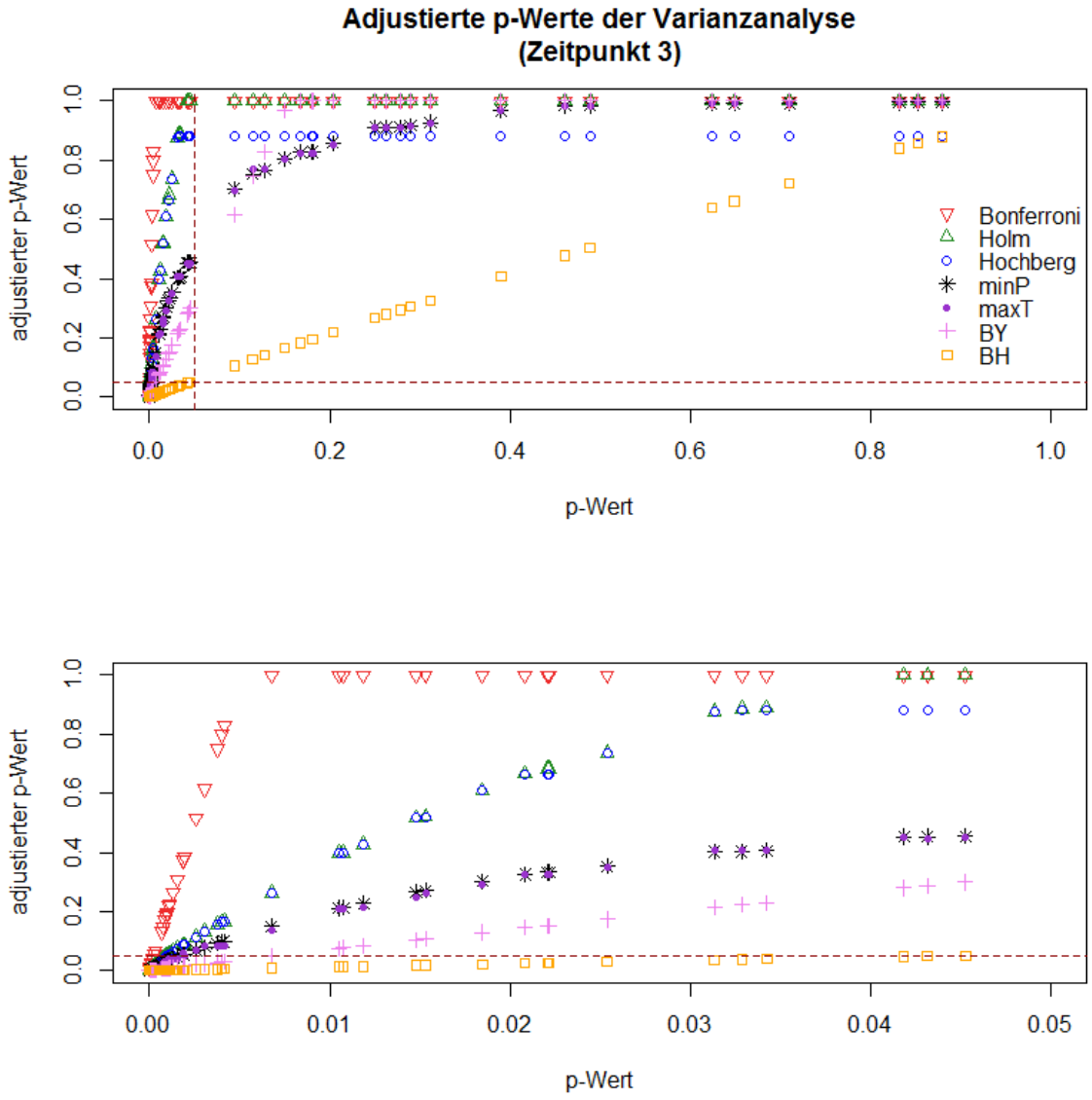

Abb. 10: *Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Die rot gestrichelten Linien sind bei 0.05 eingezeichnet um signifikante Werte erkenntlich zu machen. Oben sind alle p-Werte der Varianzanalyse abgebildet und unten nur die p-Werte der Varianzanalyse bis 0.05.*

Vergleicht man die inhaltlichen Ergebnisse, also welche Moleküle signifikante Hypothesen haben und welche nicht, kommt man auch hier zu vergleichbaren Ergebnissen. In Tabelle 11 sind die Moleküle aufgelistet, die bei allen Verfahren und zu allen Messzeitpunkten einen signifikanten p-Wert haben. Das trifft bei der Analyse mit der Varianzanalyse auf 130 Moleküle zu. Um den Vergleich mit den Ergebnissen des Kruskal-Wallis-Tests zu erleichtern, sind in dieser Tabelle die Moleküle, die dazukommen (13) also, die die nur bei der Varianzanalyse durchweg signifikante Resultate haben, mit ** markiert und die, die nur beim Kruskal-Wallis-Test ausnahmslos signifikante Ergebnisse haben, bei der Varianzanalyse aber nicht (8), mit *.

Namen der Moleküle mit "0 0 0 0" bei allen Verfahren

CH4.	X17.*	X20.	X21.	Acetylene.	Methanol.	O2.33..
X35.	X38.	X39.	ACN.	X42.	X43.	X45.
Formic.Acid.	X48.**	X49.**	X50.	X51.	X52.	X53.
X54.	X55.	X56.	X57.	X58.	X61.	X62.
X63.	SO2.	X65.	X66.	X67.	X72.	X73.
X75.**	X76.	X77.	Benzene.Xe.	X80.	X81.	X84.**
X86.	X87.	X88.	X89.*	X90.	X91.	X92.
X93.	X94.	X96.**	X98.**	X99.**	X100.	X101.
X102.	X103.	X104.	X107.	X108.	X109.	X110.
X111.**	X112.	X113.	X114.	X115.	X116.	X117.
X118.	X119.	X120.	X121.	X122.	EI.H2....M1**	EI.H2
EI.H2O.18**	EI.N2.28**	EI.O2.32**	NH3	M19	Ethylene	M29
NO	CH3NH2	M33	H2S	M36	M37	M38
M40	M41	M43	Acetaldehyde	Ethanol	M48*	M50*
M51	M52**	Butadiene	M55	M56	M57	M61
M62*	M63	M64	M67	Isoprene*	M69*	M73
M74	M75	M76	M80	M81	M87*	M90
M91	Toluene	M93	M94	M98	M101	M103
M108	M109	M115	M116	M117	M118	M119
M120	M121	M122	M123	M135		

Tabelle 11: *Auflistung aller Moleküle, die über alle Zeitpunkte und bei allen Methoden immer signifikant sind; * bedeutet nur beim Kruskal-Wallis-Test immer abgelehnt, ** bedeutet nur bei der Varianzanalyse immer abgelehnt.*

Beim Kruskal-Wallis-Test sind neun Moleküle zu beobachten, deren Nullhypothesen nie abgelehnt werden. Hier kommt zu diesen neun Molekülen eines dazu M83, wobei auch die Analyse mit dem Kruskal-Wallis-Test bei diesem Molekül fast nur nicht-signifikante Ergebnisse liefert, bis auf die Werte von Benjamini&Hochberg und des H-Tests zum Messzeitpunkt 0. Insgesamt sind es somit bei der Varianzanalyse 10 Moleküle, die kein signifikantes Resultat ergeben.

Hier bleiben folglich 60 Moleküle übrig, die je nach Messzeitpunkt und Methode mal ein signifikantes, mal ein nicht-signifikantes Resultat haben. Diese Ergebnisse ähneln, denen des Kruskal-Wallis-Tests, sind aber nicht ganz gleich. Die Tabelle 12 stellt die Ergebnisse dar, wobei die acht Moleküle, die beim Kruskal-Wallis-Test nur signifikante p-Werte hatten, hier wieder mit * markiert sind. Man erkennt, dass diese auch bei der Varianzanalyse nur wenige nicht-signifikante Ergebnisse haben.

Molekül	**Bonferroni**	**Holm**	**Hochberg**	**minP**	**maxT**	**BY**	**BH**	***H*-Test**
X17.*	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M27.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 1 1	0 0 1 1
M29..	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
Formaldehyde.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 0 0	0 0 0 0
X40.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	0 0 0 0	0 0 0 0
N2O.	1 1 1 0	1 0 1 0	1 0 1 0	1 0 1 0	1 0 1 0	1 0 0 0	0 0 0 0	0 0 0 0
HNO2.	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1
Acetic.Acid.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0
X68.	0 1 1 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1
X69.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
X70.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 1 1
X71.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0
X74.	1 0 1 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X79.	1 1 1 1	1 1 1 1	1 1 1 1	0 0 1 1	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0
X82.	0 1 1 1	0 0 1 1	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X83.	1 1 1 1	0 0 1 1	0 0 1 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0
X85.	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 0 1 1	0 0 1 0
X89.*	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X95.	0 1 1 1	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 0 0 0	0 0 0 0
X97.	0 1 1 1	0 0 1 1	0 0 1 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
X105.	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
X106.	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 1 1 0	0 1 0 0	0 0 0 0	0 0 0 0
EI.CO2.44	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M35	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 0 1 1	0 0 1 1
M39	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0	1 1 1 0	1 0 0 0	1 0 0 0
Propene	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 0 1 0	1 0 1 0
M46	1 1 1 1	1 0 1 0	1 0 1 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M47	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0	1 1 1 0
M48*	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0
M50*	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
M53	0 1 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
Acetone	1 1 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0
M62*	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0
M65	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
M66	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0

Molekül	Bonferroni	Holm	Hochberg	minP	maxT	BY	BH	H-Test
Isoprene*	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
M69*	1 1 1 1	1 0 1 1	1 0 1 1	1 0 1 1	1 0 1 1	1 0 1 0	1 0 0 0	1 0 0 0
M70	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0
M71	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 0 1 0	1 0 1 0
M72	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 1 0
M77	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
Benzene	1 1 1 1	1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 1	1 0 0 0	1 0 0 0
M79	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1	0 1 0 1	0 0 0 1	0 0 0 1
M84	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 0	0 1 1 0
M86	1 1 1 1	1 1 1 1	1 1 1 1	1 1 0 1	1 1 0 1	1 0 0 1	1 0 0 0	1 0 0 0
M87*	1 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M89	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 0 1 1	1 0 0 1	1 0 0 1
M95	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	1 1 0 0
M97	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0
M99	0 0 1 1	0 0 1 0	0 0 1 0	0 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M100	1 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M102	1 0 1 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M104	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
M105	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1
M106	1 1 1 1	1 0 0 1	1 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1	0 0 0 1
M107	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 0	0 1 1 0	0 1 1 0
M110	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
M112	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	1 1 1 1	0 1 1 1	0 1 1 1
M113	1 1 1 1	1 0 1 1	1 0 1 1	1 0 1 1	1 0 1 1	1 0 0 1	0 0 0 0	0 0 0 0
M114	1 0 0 1	1 0 0 1	1 0 0 1	0 0 0 1	0 0 0 1	0 0 0 0	0 0 0 0	0 0 0 0

Tabelle 12: *Zusammenfassung aller Ergebnisse der Varianzanalyse bzw. Anova, die je nach Zeitpunkte und Methode signifikant oder nicht signifikant sind; 0 steht für signifikant und 1 steht für nicht signifikant. Mit * markierte Moleküle haben beim Kruskal-Wallis-Test nur signifikante p-Werte.*

Im Großen und Ganzen gehen die Ergebnisse der beiden Tests auf Lageparameter in die gleichen Richtung und ähneln sich sehr, obwohl die Voraussetzungen für die Varianzanalyse bei diesem Datenbeispiel nicht optimal sind.

5.3 R-Befehle

Die eben beschriebene Analyse wurde mit R und folgenden Befehlen durchgeführt:
Die Methode MTP aus dem Packet multtest berechnet die adjustierten p-Werte nach den step-down Verfahren minP und maxT. Nachdem es sich bei diesen beiden Verfahren um Resampling-Verfahren handelt, muss diese Methode sowohl die Permutationen machen und für diese die p-Werte berechnen, als auch die nicht-adjustierten p-Werte

der Originalstichprobe. Die anderen Adjustierungsverfahren können mit der Methode p.adjust aus dem Packet stats angewendet werden. Dieser Methode werden die bereits berechneten nicht-adjustierten p-Werte übergeben und sie adjustiert diese dann mit dem gewünschten Verfahren. Damit alle Verfahren auf der gleichen Basis arbeiten können, wurden die von der Methode MTP berechneten nicht-adjustierten p-Werte auch an die Methode p.adjust zur Adjustierung der p-Werte mit den restlichen Verfahren weitergegeben. Der Befehl der Methode MTP sieht so aus $MTP(X, Y, robust, test = "f", B = 100000, method = "", nulldist = "perm", seed = 30)$. X ist die Matrix mit den Daten, wobei für jede Zeile eine Hypothese getestet wird und Y ist der Vektor mit den Gruppenbezeichnungen. Mit method = "sd.minP" oder "sd.maxT" wählt man die step-down Varianten der minP und der maxT Prozeduren. test="f" steht für die Wahl eines F-Tests, wobei mit robust=TRUE die nicht-parametrische Variante, also der Kruskal-Wallis-Test, gewählt wird und mit robust=FALSE der parametrische Test, also die Varianzanalyse. nulldist="perm" gibt an, dass eine Permutation als Resampling-Variante verwendet werden soll. B gibt an wie viele Permutationen gemacht werden sollen und mit seed wird der Startpunkt festgelegt, so dass immer dieselben Permutationen gezogen werden.
Die Adjustierungen nach Bonferroni, Holm, Hochberg, Benjamini&Yekutieli und Benjamini&Hochberg können alle mit dem Befehl $p.adjust(Z, method = "")$ ausgeführt werden, wobei method="bonferroni", "holm", "hochberg", "by" oder "bh" je nach gewünschter Methode ist und Z ist der Vektor der nicht-adjustierten p-Werte.
Für die Analyse der Verteilungen wurden folgende Befehle verwendet:
Der Shapiro-Wilk-Test wird mit $shapiro.test(Variable)$ berechnet. Der Levene-Test hat den Befehl $levene.test(Variable, Gruppenvariable)$ aus dem Packet lawstat. Für den Quartilskoeffizienten der Schiefe wurden erst die 25%-, 50%- und 75%-Quantile berechnet mit q <- $quantile(Variable, c(0.25, 0.50, 0.75), type = 1)$ und dann der Quartilskoeffizient mit $((q[3] - q[2]) - (q[2] - q[1]))/(q[3] - q[1])$. Das Wölbungsmaß von Fisher kann mit Hilfe des Befehls $kurtosis(Variable)$ aus dem Packet e1071 berechnet werden. Und um die Ausreißer zu ermitteln wurden erst die obere und die untere Grenze des Zauns berechnet mit $q[1] - 1.5 * (q[3] - q[1])$ und $q[3] + 1.5 * (q[3] - q[1])$. Und dann für jedes Molekül ermittelt wie viele Werte außerhalb dieser Grenzen liegen.

6 Fazit

Ganz allgemein ist es wohl vorteilhaft, wenn man um multiples Testen nicht herum kommt, sich im Vorfeld zu überlegen bei welchen Hypothesen es wirklich sinnvoll ist sie zu testen. Sprich nicht unbedingt notwendige Nullhypothesen erst gar nicht in die Analyse aufzunehmen, um für eine kleinere Anzahl an Hypothesen adjustieren zu müssen. So sollte man bei weiterführenden Analysen, die zum Beispiel untersuchen könnten welche Gruppen sich anhand der Messwerte der Moleküle genau voneinander

unterscheiden lassen, jene Moleküle weglassen, die bei keiner Methode und zu keinem Zeitpunkt einen signifikanten Unterschied zwischen den Erregertypen angezeigt haben. Möchte man auf jeden Fall falsch positive Ergebnisse vermeiden, sollte man eher eine $FWER$-kontrollierende Prozedur wählen, da diese konservativer sind. Am extremsten ist das Verfahren von Bonferroni, wobei dieses oft als zu konservativ gehalten wird, weshalb die schrittweisen Verfahren von Holm und Hochberg, die mehr Power haben, bevorzugt werden. Sollen mögliche Abhängigkeitsstrukturen der Variablen berücksichtigt werden, empfehlen sich die Resampling-Verfahren minP und maxT von Westfall&Young. So bieten sich diese Prozeduren bei diesem Datenbeispiel an, da es durchaus möglich ist, dass das Vorhandensein von gewissen Molekülen das anderer Moleküle positiv oder negativ beeinflusst. Sind ein paar Fehler 1. Art jedoch tolerierbar, so kann auch ein FDR-kontrollierendes Verfahren angewendet werden. Je nachdem ob die Teststatistiken unabhängig sind oder nicht, ist das Verfahren von Bejamini&Hochberg oder das von Benjamini&Yekutiele ratsam.
Bei diesem Beispiel möchte man, wie bereits erwähnt, vermutlich noch wissen zwischen welchen Gruppen sich denn hier die signifikanten Unterschiede befinden bzw. welche sich nicht unterscheiden. Denn bisher können nur Aussagen darüber getroffen werden, ob generell ein signifikanter Unterschied besteht, jedoch nicht ob dieser nur zwei Gruppen betrifft und welche das sind oder ob sich gar alle Gruppen differenzieren lassen. Um nicht wieder mit einer so großen Anzahl an Hypothesen konfrontiert zu sein, sollten bei dieser Analyse die Moleküle ausgeschlossen werden, bei denen keine Methode ein signifikantes Ergebnis gefunden hat. Da immerhin noch 125 Moleküle durchweg signifikante p-Werte haben, könnte man sich mit diesen Molekülen begnügen, wenn ein Verlust von ein paar weiteren relevanten Molekülen verkraftbar ist. Sonst muss man sich überlegen welche Adjustierungsmethode als Maß dienen soll. Die von Benjamini&Hochberg zum Beispiel lehnt kaum weniger Hypothesen ab, als der Kruskal-Wallis-Test ganz ohne Adjustierungen, ist also möglicherweise nicht strikt genug und verbirgt einige falsch Positive. Außerdem dürfte interessant sein, ob auch die einzelnen Erreger innerhalb einer Gruppe mit Hilfe dieser Messungen voneinander unterschieden werden können.
Und schließlich genügt es nicht sich nur Gedanken über die Art der Adjustierung zu machen, sondern das grundlegende Analyseverfahren muss ebenfalls richtig gewählt sein. Denn wenn schon die nicht-adjustierten p-Werte falsch berechnet wurden, bringt die Adjustierung auch nicht mehr viel. Bei diesem Beispiel scheint die Verletzung der Normalverteilungsannahme und der Homoskedastizität keine große Auswirkung zu haben. Trotzdem sollte man sich immer bewusst sein, ob denn die Voraussetzungen erfüllt werden oder nicht und im Zweifelsfall auch eine nicht-parametrische Alternative anwenden und dann die Ergebnisse vergleichen.

7 Literaturverzeichnis

Benjamini, Y. / Hochberg, Y. (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. - In: Journal of the Royal Statistical Society, Series B, 57(1), 289–300.

Benjamini, Y. / Yekutieli, D. (2001): The control of the false discovery rate in multiple testing under dependency. - In: Annals of Statistics, 29(4), 1165–1188.

Dudoit, Sandrine / Shaffer, Juliet Popper / Boldrick, Jennifer C. (2003): Multiple Hypothesis Testing in Microarray Experiments. - In: Statistical Science, 2003, Vol. 18, No. 1, 71-103, S. 73-74 und S. 78-81.

Fahrmeir, Ludwig / Kunstler, Rita / Pigeot, Iris /Tutz, Gerhard (2010): Statistik: Der Weg zur Datenanalyse, 7.Auflage, Berlin (Springer Verlag), S. 48, S. 66f., S. 74-76, S. 428, S. 516-519 und S. 527f.

Hochberg, Y. (1988): A sharper Bonferroni procedure for multiple tests of significance. - In: Biometrika, 75(4), 800–802.

Holm, S. (1979): A simple sequentially rejective multiple test procedure. - In: Scandinavian Journal of Statistics, 6, 65–70.

Sachs, Lothar / Hedderich, Jürgen (2009): Angewandte Statistik. Methodensammlung mit R. 13. Auflage, Berlin Heidelberg (Springer Verlag), S. 361, S. 397f, S. 489-491, S. 498f und S. 514f.

Shapiro, S. S. / Wilk M. B. (1965): An Analysis of Variance Test for Normality (Complete Samples). - In: Biometrika, 52(3-4), 591-611, S. 592f.

Westfall, P. H. / Young, S. S. (Hg.) (1993): Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York.

Zierer, Astrid (2013): Multiples Testen und die Kontrolle der gFWER in der statistischen Analyse genetischer Daten. 1. Auflage, München (Verlag Dr. Hut), S. 20, S. 32, S. 35, S. 37, S. 42f, S. 46-48, S. 50-52 und S. 54f.

Eigenständigkeitserklärung:

Ich versichere, dass ich die vorgelegte Bachelorarbeit eigenständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen verwendet und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Bachelorarbeit ist in dieser oder einer ähnlichen Form in keinem anderen Kurs vorgelegt worden.

München, den