Aus dem Institut für Medizinische Biometrie und Statistik
der Universität zu Lübeck
Direktor: Prof. Dr. rer. nat. Andreas Ziegler

Zur Gewichtung von Familien nach Markerinformativität in modellfreien Kopplungsanalysen

Inauguraldissertation

zur
Erlangung der Doktorwürde
der Universität zu Lübeck
– Aus der Medizinischen Fakultät –

vorgelegt von
Daniel Franke
aus Waldshut-Tiengen

Lübeck 2006

1. Berichterstatter(in):	Prof. Dr. rer. nat. Andreas Ziegler
2. Berichterstatter(in):	Prof. Dr. rer. nat. Lutz Mattner
Tag der mündlichen Prüfung:	18.10.2006

Zum Druck genehmigt. Lübeck, den 18.10.2006

gez. Prof. Dr. med. Werner Solbach

– Dekan der medizinischen Fakultät –

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einführung

1.1 Zielsetzung

Zur Analyse genomweiter Untersuchungen erkrankter Geschwisterpaare wurde eine Vielzahl von statistischen Tests entwickelt und angewandt. Aufgrund der begrenzten Genauigkeit bzw. Auflösung der bisher üblicherweise verwendeten Mikrosatelliten können die Statistiken die potentiell vorhandene Informativität aber selbst in Mehrpunktanalysen nicht vollständig ausnutzen.

Zur vollen Nutzung der Markerinformationen wurde bereits zum Genetic Analysis Workshop (GAW) 13 ein, sich auf quantitative Phänotypen beziehenden, Gewichtungsansatz vorgeschlagen (Jacobs et al., 2003). Eine spätere Arbeit bekräftigt die Notwendigkeit eines solchen Gewichtungsansatzes nach Markerinformationen auch für Analysen binärer Phänotypen (Schork und Greenwood, 2004).

Im allgemeinen Kontext der Meta-Analysen läßt sich zeigen, daß sich die statistische Güte deutlich verbessert wenn Einzelstudien nach ihrer Informativität gewichtet werden (Loesgen et al., 2001; Dempfle und Loesgen, 2004).

Ziel dieser Arbeit ist es zu überprüfen, ob sich durch Gewichtung nach Markerinformativität auch bei klassischen Kopplungsstudien eine höhere statistische Güte bei gleichbleibender Fehlerwahrscheinlichkeit erster Art erreichen läßt. Hierzu werden in ausführlichen Simulationsstudien zwei neue Methoden evaluiert, welche aus Erweiterung bekannter Standardverfahren hervorgegangen sind. Für quantitative Phänotypen wird eine Adaption des Verfahrens von Haseman und Elston (1972) betrachtet; für binäre Phänotypen eine neue Variante der Mittelwerttests nach deVries et al. (1976) und Green und Woodrow (1977).

1.2 Aufbau der Arbeit

Der Aufbau der weiteren Arbeit ist in drei Teile gegliedert: das nachfolgende Kapitel 2 beginnt mit einer Zusammenfassung der in dieser Arbeit verwendeten Methoden. In den Kapiteln 3 und 4 werden Verfahren betrachtet, wie sowohl die etablierte Methode von Haseman und Elston (1972) zur Analyse quantitativer Phänotypen, als auch Mittelwerttests zur Analyse binärer Phänotypen (deVries et al., 1976; Green und Woodrow, 1977), mit Gewichtung nach Markerinformationen erweitert werden können. Eine Diskussion der Resultate erfolgt in Kapitel 5.

Kapitel 2 beschreibt einige, nicht allein auf diese Arbeit beschränkte, Methoden. In Abschnitt 2.1 werden verschiedene Kopplungsverfahren und ihre

Einsatzgebiete skizziert. Es werden sowohl modellfreie als auch modellbasierte Algorithmen beschrieben (Abschnitt 2.1.1), sowie auf die Unterschiede zwischen Zwei- und Mehrpunktanalysen eingegangen (Abschnitte 2.1.2 bzw. 2.1.3). Für die in den Kapiteln 3 und 4 betrachteten statistischen Methoden wird ein Maß zur Bestimmung der genetischen Ähnlichkeit zwischen Paaren von verwandten Personen benötigt. Abschnitt 2.2 stellt hierfür die notwendigen Grundlagen bereit. In Abschnitt 2.2.1 wird die Anzahl der Allele *identisch nach Herkunft* (engl.: identical by descent, IBD) als ein Maß der genetischen Ähnlichkeit vorgestellt. Auf den Anteil der Allele IBD, dem zugehörigen Verhältnismaß, wird im nachfolgenden Abschnitt 2.2.2 eingegangen. Über den Raum der IBD-Verteilungen (Abschnitt 2.2.3) werden Informativität der IBD-Werte sowie Abstandsdefinitionen im Raum der IBD-Werte hergeleitet werden (Abschnitte 2.2.4 und 2.2.5). Aus den in Abschnitt 2.2.5 dargestellten Abstandsmaßen werden in den Kapiteln 3 und 4 jeweils plausibel erscheinende Gewichtungsschemata für die jeweilige Methode abgeleitet. Um die um Gewichtung nach Markerinformation erweiterten Methoden zu validieren wurden Simulationsstudien durchgeführt. In Abschnitt 2.3 wird erläutert, wie sich statistische Test empirisch verifizieren lassen. Hierzu werden in den Abschnitten 2.3.1 und 2.3.2 Anteilsschätzer beschrieben, mit deren Hilfe sich Fehlerrate bzw. statistische Güte empirisch ermitteln lassen. Aussagen über die Präzision dieser Schätzer werden schließlich in Abschnitt 2.3.3 getroffen.

In Kapitel 3 erfolgt die Erweiterung des Verfahrens von Haseman und Elston (1972) um Gewichtung nach Markerinformativität. Hierzu wird in Abschnitt 3.1 das ursprüngliche Verfahren von Haseman und Elston skiz-

ziert. Außerdem werden einige andere, bereits bekannte, Gewichtungsvarianten vorgestellt und auf ihre Relevanz in Bezug auf diese Arbeit hin untersucht. Auf die mathematischen Grundlagen des Verfahrens von Haseman und Elston (1972) wird in Abschnitt 3.2 eingegangen. Über das lineare Modell von Haseman und Elston (1972) wird das allgemeine lineare Modell motiviert (Abschnitte 3.2.1 und 3.2.2). Die Funktionen zur Schätzung der Varianz-Kovarianzmatrix zur Gewichtung nach Markerinformativität, die im allgemeinen linearen Modell verwendet werden sollen, werden in Abschnitt 3.2.3 entwickelt. Desweiteren wird in Abschnitt 3.2.4 die Äquivalenz von Regression und Umkehrregression im Sinne des Kopplungstests gezeigt. Damit wird ein Resultat von Schaid et al. (2003, Appendix A) verallgemeinert. Zur Validierung des entwickelten Ansatzes der Gewichtung nach Markerinformativität wurden Simulationsstudien durchgeführt (Abschnitt 3.3). Die verwendeten Datenmaterialien wurden sowohl intern (Abschnitt 3.3.1) als auch extern (Abschnitt 3.3.2) simuliert. Aus den Ergebnissen der Simulationsstudien lassen sich die in Abschnitt 3.4 dargestellten Schlüsse ziehen. Es besteht die Gefahr, daß Studien durch das *Design von p-Werten* über plausibel erscheinende Gewichtungsschemata manipuliert werden könnten.

Kapitel 4 beschreibt die Erweiterung der klassischen Mittelwerttests für erkrankte Geschwisterpaare um eine Gewichtung nach Markerinformation. In Abschnitt 4.1 wird eine kurze Übersicht über vorhandene Methoden gegeben. Die Mittelwerttests in ihrer klassischen Form werden in Abschnitt 4.1.1 genauer dargestellt. Eine Übersicht über bereits vorhandene Gewichtungsansätze und ihre Relevanz erfolgt in Abschnitt 4.1.2. Die mathematischen Herleitungen, die zum nach Markerinformativität gewichteten Mittelwert-

test führen, werden in Abschnitt 4.2 gezeigt. Abschnitt 4.2.1 faßt die klassischen Mittelwerttests mit ihren Voraussetzungen und Annahmen nochmals zusammen, während in Abschnitt 4.2.2 die gewichtete Teststatistik, inklusive ihrer asymptotischen Verteilung, hergeleitet wird. Desweiteren wird ein, auf Abschnitt 2.2.5 aufbauendes, Gewichtungsschema entwickelt. Zu Beginn dieser Arbeit war nicht abzusehen, daß es möglich sein würde, die, im Abschnitt 4.2.2 hergeleitete, asymptotische Verteilung anzugeben. Daher wurde auch ein neuartiges, auf der Simulation von IBD-Werten gleicher Informativität basierendes Verfahren zur empirischen Ermittlung der p-Werte entwickelt (Abschnitt 4.2.3). Wie zuvor wurden auch für den gewichteten Mittelwerttest ausführliche Simulationsstudien durchgeführt (Abschnitt 4.3). In Abschnitt 4.4 folgen Anwendungen der gewichteten Teststatistik: zum einen auf einen hypothetischen Datensatz (Abschnitt 4.4.1), zum anderen auf einen zuvor veröffentlichten Realdatensatz (Risch, 1990; Mein et al., 1998).

Kapitel 2

Allgemeine Methoden

2.1 Testen auf Kopplung

Aufgrund der hohen Variablität des humanen Genoms und seiner, auch für heutige Begriffe immer noch, immensen Informationsdichte, ist es im Allgemeinen nicht möglich krankheitsverursachende Veränderungen durch eine direkte Suche zu finden. Stattdessen werden eine Vielzahl indirekter statistischer Methoden verwendet um Abweichungen zwischen den mendelschen Erbgängen und einer überzufällig häufigen Weitergabe von Genvariationen innerhalb von Familien aufzuspüren. Diese statistischen Methoden werden als *Kopplungsverfahren* bezeichnet, sie untersuchen die *Kopplung* eines Krankheitsbildes, eines *Phänotyps*, an einen Genort. Genauer ausgedrückt untersuchen Kopplungsverfahren die Kopplung zwischen zwei Genorten, einem bekannten *Markergenort* und einem unbekannten Locus, der die Krankheit bedingt. Je örtlich näher der untersuchte Markergenort am putativen Krankheitsgenort liegt, desto stärker ist deren Kopplung.

Kopplungstests nutzen einen einfachen Zusammenhang aus: benachbarte Genorte können in der Meiose durch Rekombinationen getrennt werden. Sind die betrachteten Loci ungekoppelt so ist das Auftreten bzw. das Ausbleiben einer Rekombination zwischen ihnen gleichwahrscheinlich. Sind sie gekoppelt, so sinkt die Wahrscheinlichkeit von Rekombinationen – je stärker die Kopplung, desto geringer die Wahrscheinlichkeit einer Rekombination.

2.1.1 Modellfreie und modellbasierte Methoden

Modellfreie Methoden kommen zum Einsatz, wenn keinerlei Gesetzmäßigkeiten (Mendelsche Vererbungsmuster) für die Vererbung des Phänotyps bekannt sind. Sie versuchen die Frage zu klären, ob die erkrankten Personen eines Stammbaumes an einem bestimmten Genort mehr Allele gemeinsam haben als ohne Kopplung zu erwarten wäre. Hierbei ist es wichtig, zwischen den gebräuchlichen Maßen der genetischen Ähnlichkeit zu unterscheiden: Es ist von erheblicher Bedeutung, ob Allele nur *identisch nach Zustand* (engl.: identical by state, IBS) oder *identisch nach Herkunft* (engl.: identical by descent, IBD) sind. Ein überzufällig hoher Anteil Allele IBD unter den Kranken deutet auf den gesuchten Zusammenhang zwischen Genort und Krankheit hin.

Modellbasierte kopplungsanalytische Verfahren legen die Annahme eines exakt bekannten mendelschen Vererbungsmusters zugrunde, welches beispielsweise durch eine vorgeschobene *Segregationsanalyse* bestimmt werden kann. Aufgrund ihrer Spezifikation sind modellbasierte Verfahren auf große Stammbäume mit mehreren erkrankten Individuen zugeschnitten. Ihr Haupt-

einsatzgebiet liegt in monogenen oder komplexen Erkrankungen, die einem mendelschen Vererbungsmuster folgen (vgl. Ott, 1999; Strauch, 2002).

Modellfreie Verfahren sind, sofern das Vererbungsmodell bekannt ist, hinsichtlich ihrer Güte und Trennschärfe den modellbasierten Test unterlegen, andernfalls verhalten sie sich robuster als ihre klassischen Pendants bei mißspezifiziertem Vererbungsmuster (Ziegler und König, 2006).

2.1.2 Zweipunktanalysen

Sollen jeweils exakt zwei Genorte daraufhin überprüft werden, ob Kopplung zwischen ihnen besteht, spricht man von *Zweipunkt-* oder *Twopoint-Analysen*. Modellbasierte Analysen schätzen mittels Zweipunktanalysen die Rekombinationsfrequenz zwischen den Genorten – dabei ist es unerheblich ob die Frequenz zwischen zwei bereits bekannten Markern oder einem Marker und einem noch unbekannten potentiellen Krankheitsgenort geschätzt werden soll. Ersteres wird zur Erstellung von Marker-Karten eines Chromosomes, letzteres zur Lokalisation des Krankheitsgenorts innerhalb eines Chromosoms verwendet. Bei komplexen Krankheiten, bei denen mehrere Genorte für einen (Krankheits-) Phänotyp verantwortlich sind, greift der Zweipunkt-Ansatz zu kurz, die Trennschärfe läßt nach (Ziegler und König, 2006, Kap. 6). In einem solchen Fall sind Mehrpunktanalysen vorzuziehen.

2.1.3 Mehrpunktanalysen

Werden simultan mehr als zwei Genorte in eine Auswertung einbezogen spricht man von *Mehrpunkt-* oder *Multi-Marker-Analysen*. Statt wie bei den

Zweipunktansätzen den unbekannten Krankheitsgenort gegen einen einzelnen bekannten Marker zu testen, wird bei Mehrpunkanalysen gegen eine ganze Gruppe von Markern gekopppelt, deren Abstände und Reihenfolge untereinander genau bekannt sein müssen. Die Mehrpunkt- hat gegenüber der Zweipunktanalyse den Vorteil des *übertragens von Informationen* (engl.: spill-over), d.h. informativere Genorte geben an benachbarte, uninformativere, Marker weiter. Dies ist besonders bei der Verwendung von Einzelnukleotidpolymorphismus (engl.: Single Nucleotide Polymorphisms, SNPs) von großer Bedeutung, da deren Einzelinformativität nur gering ist.

2.2 Maße der genetischen Ähnlichkeit und ihre Informativität

Die in dieser Arbeit betrachteten statistischen Methoden, sowohl das Verfahren von Haseman und Elston (Kapitel 3) als auch die Mittelwerttests für erkrankte Geschwisterpaare (Kapitel 4) benötigen ein Maß zur Bestimmung der genetischen Ähnlichkeit zwischen Paaren von Personen. In diesem Abschnitt soll ein solches Maß zur Bestimmung der genetischen Ähnlichkeit ausführlich vorgestellt werden (Abschnitte 2.2.1 und 2.2.2). Auf den von diesem Maß aufgespannten Raum wird in Abschnitt 2.2.3 genauso eingegangen werden wie auf Informations- und Abstandsmaße (Abschnitte 2.2.4 bzw. 2.2.5).

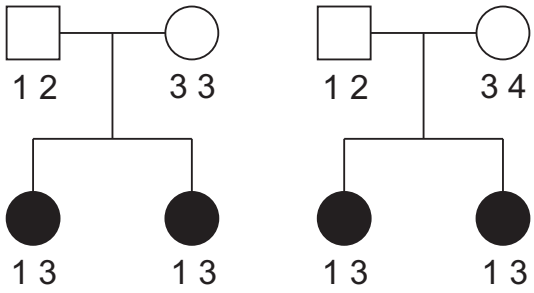


Abbildung 2.1: Mating-types, exemplarisch

2.2.1 Der IBD-Wert

Zur Messung genetischer Ähnlichkeit zwischen zwei Personen können verschiedene Maße verwendet werden. Im allgemeinen werden Ähnlichkeitsmaße über die Anzahl der identischen Allele an einem Genort bestimmt. Hierbei muß nach der Anzahl der Allele *identisch nach Zustand* (engl.: identical by state, IBS) und *identisch nach Herkunft* (engl.: identical by descent, IBD) unterschieden werden. Die Anzahl der Allele IBS haben als Maß der genetischen Ähnlichkeit in der Praxis nur eine geringe Bedeutung, so daß auf eine weiterführende Betrachtung hier verzichtet wird (vgl. Bishop und Williamson, 1990).

Abbildung Abb. 2.1 zeigt exemplarisch zwei der sieben möglichen Paarungsmöglichkeiten (engl.: mating-types) elterlicher Markerallele für ein Geschwisterpaar. Die Nachkommen des linken Elternpaares teilen sich zwar beide Allele gemeinsam nach Zustand, aber es ist nicht klar, welches großelterliche Allel von der Mutter vererbt wurde. Die Anzahl der Allele identisch nach Herkunft ist nicht eindeutig. Die Wahrscheinlichkeiten, daß sich die Kinder exakt ein Allel IBD teilen, ist genau so groß wie die Wahrscheinlichkeit dafür, daß sie sich zwei Allele teilen: $P\left(IBD = 1\right) = \frac{1}{2}$ und $P\left(IBD = 2\right) = \frac{1}{2}$. Im rechten Teilbild, mit der heterozygoten Mutter ist die Situation eindeutig. Die

Kinder teilen sich zwei Allele IBD. Hierbei ist zu beachten, daß IBD-Werte nur zwischen verwandten Personen mit gemeinsamen Vorfahren bestimmt werden können.

Eine *IBD-Schätzung* $z = (z_0, z_1, z_2)$ wird durch ein Wahrscheinlichkeitstripel dargestellt. Für beliebige IBD-Tripel gilt: $z_0 + z_1 + z_2 = 1$. Hierbei geben die Wahrscheinlichkeiten z_i an, wie wahrscheinlich es ist, daß sich ein Paar an einem Genort genau i Allele IBD teilt: $z_i = P(IBD = i)$. Das linke Geschwisterpaar in Abb. 2.1 stellt eine IBD-Schätzung von $\left(0, \frac{1}{2}, \frac{1}{2}\right)$ dar, das rechte dagegen die eindeutigen Wahrscheinlichkeiten $(0, 0, 1)$.

2.2.2 Der Anteil Allele IBD

Häufig werden in Kopplungsanalysen nicht die Wahrscheinlichkeiten, 0, 1 oder 2 Allele IBD zu haben, direkt verwendet, sondern es wird ein komprimiertes Maß herangezogen, *der Anteil der Allele IBD*. Der Anteil Allele IBD τ eines IBD-Werts $z = (z_0, z_1, z_2)$ ist definiert durch:

$$\tau = \frac{2z_2 + z_1}{2} \tag{2.1}$$

Mit dem Anteil Allele IBD wird die genetische Ähnlichkeit prägnant als Verhältnis ausgedrückt: sind die untersuchten Personen an einem Genort unähnlich, so liegt der Anteil der Allele IBD bei 0, sind sie sich genetisch ähnlich liegt er bei 1.

Desweiteren kann der Anteil Allele IBD als die Wahrscheinlichkeit, daß zwei Personen an einem Marker exakt ein Allel IBD haben, angesehen werden.

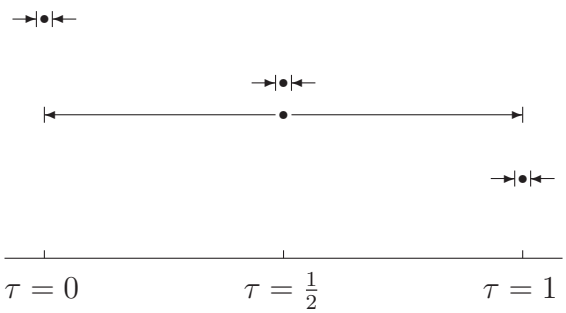


Abbildung 2.2: Anteile der Allele IBD für verschiedene IBD-Werte: während der Anteil der Allele IBD für $(1, 0, 0)$, $(0, 1, 0)$ und $(0, 0, 1)$ punktgenau festgelegt ist, „verschmiert“ eine Schätzung von $(\frac{1}{2}, 0, \frac{1}{2})$ über den gesamten Bereich. Der Anteil Allele IBD $\tau = \frac{1}{2}$, ist aber zu gleichen Teilen entweder 0 oder 1.

Festzuhalten bleiben zwei Eigenarten des Anteils der Allele IBD: zum einen ist der Anteil der Allele IBD als Maß der genetischen (Un-)Ähnlichkeit verlustbehaftet, d.h. nur mit der Angabe des Anteils ist es im Allgemeinen nicht möglich auf den ursprünglichen IBD-Wert zurückzuschließen. Während die Anteile der Allele IBD für IBD-Werte von $(1, 0, 0)$ bzw. $(0, 0, 1)$ mit $\tau = 0$ bzw. $\tau = 1$ eindeutig sind, ist bereits ein IBD-Wert von $(0, 1, 0)$ mit $\tau = \frac{1}{2}$ mehrdeutig: alle IBD-Werte der Form $\left(\frac{1-a}{2}, a, \frac{1-a}{2}\right)$, für $0 \leq a \leq 1$, ergeben einen Anteil Allele IBD von $\frac{1}{2}$. Zum anderen ist der Anteil der Allele IBD eine eher ungenaue Abbildung der zugehörigen IBD-Werte. Abbildung 2.2 zeigt dies exemplarisch: der zum IBD-Wert $(\frac{1}{2}, 0, \frac{1}{2})$ gehörende Anteil Allele IBD beträgt $\frac{1}{2}$, der wahre Wert ist mit gleichen Wahrscheinlichkeiten entweder 0 oder 1.

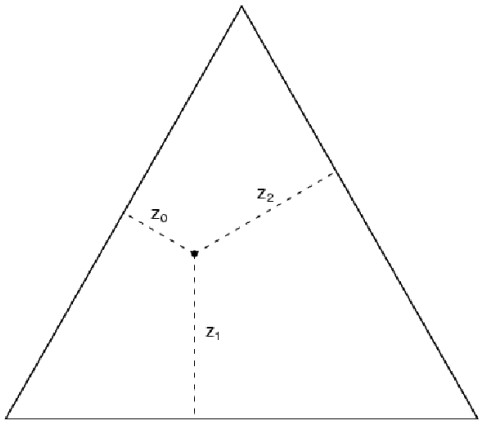


Abbildung 2.3: Diagramm nach de Finetti, der Raum der IBD-Verteilungen. Die Summe der Höhen in jedem Punkt entspricht der Höhe des Dreiecks (Satz von Viviani). Wird die Höhe des Dreiecks zu 1 gewählt entspricht jeder Punkt einem gültigen IBD-Wert mit $z_0 + z_1 + z_2 = 1$.

2.2.3 Der Raum der IBD-Verteilungen

Wie bereits Li (1955, S. 8, Abb. 2) gezeigt hat, lässt sich der eigentlich dreidimensionale Raum der Genotypverteilungen einfach und anschaulich in einem zweidimensionalen Diagramm nach de Finetti darstellen (Abb. 2.3).

Nach dem Satz von Viviani ist in einem gleichseitigen Dreieck die Summe der Abstände eines Punktes zu den Seiten konstant und entspricht der Höhe des Dreiecks. Wird daher die Höhe des gleichseitigen Dreiecks zu 1 gewählt und die Lote jeden Punktes z auf die Seiten des Dreiecks mit z_0, z_1 respektive z_2 bezeichnet (Abb. 2.3), so entspricht jeder Punkt innerhalb als auch auf den Kanten des Dreiecks, einer gültigen IBD-Verteilung mit $z_0 + z_1 + z_2 = 1$.

2.2.4 Informativität von IBD-Werten

Da sich die in der weiteren Arbeit betrachteten Methoden auf Geschwisterpaare beschränken, ist dieser Abschnitt auch in diesem Kontext zu sehen. Informativität, wie sie hier für Geschwisterpaare definiert werden wird, ließe sich auch auf beliebige Verwandtschaftsrelationen verallgemeinern.

Geschwisterpaare, deren Allele an einem Genort keinerlei Rückschlüsse auf ihre Vererbung zulassen, bringen keinerlei zusätzliche Information in die Analyse dieses Genorts ein – sie sind uninformativ. Die IBD-Verteilung u an einem uninformativen Genort entspricht $(u_0, u_1, u_2) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$. Weicht eine IBD-Schätzung von diesem uninformativen Wert ab, so wird sie als informativ angesehen. Je größer die Abweichung, desto größer die Informativität (vgl. Kruglyak und Lander, 1995). Kann eine IBD-Verteilung eindeutig bestimmt werden, so ist ihre Informativität maximal.

Das de Finetti Diagramm in Abbildung 2.3 illustriert den Begriff der Abweichung sehr anschaulich: eindeutige IBD-Werte wie $(1, 0, 0)$, $(0, 1, 0)$ oder $(0, 0, 1)$ markieren die Eckpunkte des Dreiecks, die Unsicher- und Unbestimmtheiten nehmen im inneren Bereich zu.

Bisher wird in den meisten auf Geschwisterpaaren basierenden Analysen die Informativität der IBD-Schätzer nicht berücksichtigt. Das heißt, die aus dem Anteil Allele IBD resultierenden Mehrdeutigkeiten (vgl. Abschnitt 2.2.2) werden außer Acht gelassen. Eine der wenigen Ausnahmen bildet GENEHUNTER (Kruglyak und Lander, 1995). Die Software läßt nur Paare in ihre Analysen einfließen, die nicht vollständig uninformativ sind.

Kapitel 3 und 4 zeigen Wege auf, wie man die Informativität der IBD-Schätzer in verschiedene Methoden einbeziehen kann. In den nachfolgenden Abschnitten werden hierfür die Grundlagen gelegt. Es werden verschiedene Maße zur Bestimmung von Entfernungen zwischen IBD-Werten betrachtet, insbesondere der Abstand eines beliebigen IBD-Wertes zur uninformativen a-priori Schätzung $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$.

2.2.5 Abstandmaße bezüglich der IBD-Informativität

Intuiutiv läßt sich sagen, daß je weiter eine IBD-Schätzung von $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ entfernt ist, desto höher ihre Informativität ist, desto größer sollte ihr Gewicht innerhalb einer Analyse sein. Die formale Festlegung von „weiter entfernt“ ist hierbei nicht trivial. Eine Abstandsdefinition, die für eine Anwendung optimal ist, könnte in einer anderen Anwendung unbrauchbar sein. Entsprechend sind die hier vorgestellten Varianten der Gewichtung nach Informativität einzig unter dem Gesichtspunkt der Plausibilität zu sehen. Anwendbarkeits- und Optimalitätskriterien werden in den jeweiligen Anwendungskapiteln betrachtet.

Euklidische Distanz

Die euklidische Distanz im Raum der IBD-Verteilungen wurde bereits vorgestellt (Franke et al., 2005), soll hier aber nochmals vollständig und ausführlich hergeleitet werden.

Um in der x-y-Ebene des de Finetti Diagramms (Abb. 2.3) den euklidischen Abstand d zweier IBD-Schätzer v und w zu bestimmen, müssen die dreidimensionalen Koordinaten $v = (v_0, v_1, v_2)$ bzw. $w = (w_0, w_1, w_2)$ auf ihre zweidimensionalen Pendants, $v' = (v_x, v_y)$ bzw. $w' = (w_x, w_y)$, abgebildet werden. Mittels der Definition des euklidschen Abstandes im $\mathbb{R}^2$ $d(v', w') = \sqrt{(v_x - w_x)^2 + (v_y - w_y)^2}$ wird dann die Entfernung d zwischen v' und w' bestimmt.

Die vektorwertige Transformationsfunktion S, welche einen IBD-Wert $v = (v_0, v_1, v_2)$ auf die kartesischen Koordinaten (v_x, v_y) abbildet, läßt sich

durch Auflösung des nachfolgenden, überbestimmten Gleichungssystems finden.

$$\begin{aligned} v_y &= \sqrt{3}\, v_x - 2\, v_0 \\ v_y &= v_1 \\ v_y &= -\sqrt{3}\, v_x + 2\,(1 - v_1) \end{aligned}$$

Durch Gleichsetzen der ersten und der dritten Gleichung ergibt sich:

$$S\,(v_0, v_1, v_2) = (v_x, v_y) = \left(\frac{2\, v_0 + v_1}{\sqrt{3}}, v_1 \right) \quad .$$

Somit ist der allgemeine euklidische Abstand d zweier beliebiger IBD-Schätzer v und w in der Ebene des de Finetti Diagramms durch

$$d_{euklid}\,(S(v), S(w)) = \sqrt{\left(\frac{2v_0 + v_1}{\sqrt{3}} - \frac{2w_0 + w_1}{\sqrt{3}} \right)^2 + (v_1 - w_1)^2}$$

gegeben. Im Speziellen gilt besonders für den Abstand einer IBD-Schätzung v zum a-priori IBD-Wert $u = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ eines Geschwisterpaares :

$$d_{euklid}\left(S\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right), S(v) \right) = \sqrt{\frac{(v_0 - v_2)^2}{3} + \frac{(1 - 2v_1)^2}{4}} \quad . \tag{2.2}$$

IBD-Norm

Eine Distanz kann nicht nur für die x-y-Ebene des de Finetti Diagrams, sondern auch für den dreidimensionalen Raum der IBD-Verteilungen definiert werden. Hierzu wird der IBD-Wert als dreidimensionaler Vektor betrachtet,

der den bekannten Vektoroperationen, insbesondere der Norm-Bildung, unterworfen werden kann.

Die p-Norm des Differenz-Vektors zwischen den IBD-Werten v und w ist definiert als

$$d_{p-norm}(v,w) = \sqrt[p]{(v_0 - w_0)^p + (v_1 - w_1)^p + (v_2 - w_2)^p}$$

wobei die 2-Norm ($p = 2$) dem euklidischen Abstand der Punkte v und w im dreidimensionalen entspricht. Es ist zu beachten, daß der Differenzvektor $v - w$ nicht notwendigerweise eine gültige IBD-Verteilung darstellt. Seine Länge repräsentiert den Abstand zwischen v und w, keine IBD-Verteilung. Als Beispiel seien hier $v = (1,0,0)$ und $w = (0,0,1)$ angeführt: $v - w = (1,0,-1)$. Je nach Wahl des Parameters p ergeben sich verschiedene Abstände zwischen v und w. In diesem Beispiel beträgt der Abstand $\sqrt[p]{2}$ für gerade p bzw. 0 für ungerade p.

Wie zuvor gilt im Speziellen besonders der Abstand einer IBD-Schätzung v zur uninformativen a-priori IBD-Schätzung u, $u = (u_0, u_1, u_2) = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$:

$$d_{p-norm}\left(\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right), v\right) = \sqrt[p]{\left(\frac{1}{4} - v_0\right)^p + \left(\frac{1}{2} - v_1\right)^p + \left(\frac{1}{4} - v_2\right)^p} \quad . \tag{2.3}$$

Prinzipiell können auch andere Definitionen einer Vektor-Norm, beispielsweise die Maximumbetragsnorm, $d_{max-norm}(v,w) = \max_{i=0,1,2} |v_i - w_i|$, zur Definition eines Gewichtes verwendet werden.

Entropie als Maß der Unähnlichkeit

Nach Ewens und Grant (2001) ist die relative Entropie H als Maß der Unähnlichkeit zwischen IBD-Schätzern $v = (v_0, v_1, v_2)$ und $w = (w_0, w_1, w_2)$ definiert als

$$H\left(v \| w\right) = \sum_{k=0}^{2} v_k \log \frac{v_k}{w_k} \quad .$$

Man beachte hierbei, daß H nicht kommutativ ist, d. h. $H(v\|w) \neq H(w\|v)$. Sei u die a-priori IBD-Schätzung bei Geschwisterpaaren mit $u = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$, so ist der normierte Abstand einer beliebigen IBD-Schätzung $v = (v_0, v_1, v_2)$ zu u nach relativer Entropie gegeben durch:

$$d_{shannon}\left(v, \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)\right) = \frac{v_0 \log(4\, v_0) + v_1 \log(2\, v_1) + v_2 \log(4\, v_2)}{\log 4} \qquad (2.4)$$

Eine modifizierte Form dieses Informationsmaßes ergibt sich, wenn statt des uninformativsten IBD-Schätzers derjenige mit der größten Entropie $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ als Bezugspunkt gewählt wird:

$$d_{modshannon}\left(v, \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)\right) = \frac{v_0 \log(3\, v_0) + v_1 \log(3\, v_1) + v_2 \log(3\, v_2)}{\log 3} \qquad (2.5)$$

Varianz

Ein anderes Informationsmaß wurde bereits von Kruglyak und Lander (1995) beschrieben. Die Informativität einer IBD-Schätzung v wird definiert durch das Verhältnis der Varianz von v zur Varianz einer uninformativen Schätzung

$u, u = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$, d. h. .

$$d_{varianz}(v) = 1 - \frac{\mathbb{V}ar\,(v)}{\mathbb{V}ar\,(u)}$$

wobei Erwartungswert und Varianz eines IBD-Wertes v gegeben sind durch

$$\begin{aligned} \mathbb{E}\,(v) &= v_1 + 2\,v_2 \\ \mathbb{E}\left(v^2\right) &= v_1 + 4\,v_2 \\ \mathbb{V}ar\,(v) &= \mathbb{E}\left(v^2\right) - (\mathbb{E}\,(v))^2 = v_1(1 - v_1) + 4\,v_0\,v_2 \end{aligned}$$

Somit ist für die IBD-Schätzung $v, v = (v_0, v_1, v_2)$, das Informationsmaß bezüglich der Varianz definiert als:

$$d_{varianz}(v) = 1 - 2v_1(1 - v_1) - 8v_0v_2 \tag{2.6}$$

Für den praktischen Einsatz ist zu beachten daß dieses Informationsmaß negativ werden kann:

$$d_{varianz}\left(\frac{1}{2}, 0, \frac{1}{2}\right) = -1$$

Auf diesen Umstand haben Kruglyak und Lander (1995) bereits hingewiesen.

2.3 Empirische Verifikation statistischer Tests mittels Simulationsstudien

Die in den Kapiteln 3 und 4 zu entwickelnden Testverfahren sollen schließlich durch Monte-Carlo-Simulationsstudien verifiziert werden. Im Folgenden

werden in diesem Kapitel Methoden zur empirischen Validierung statistischer Tests unter Verwendung von Simulationsstudien zusammengefaßt. In den Abschnitten 2.3.1 und 2.3.2 wird erläutert, wie Fehlerhäufigkeiten und statistische Güte mittels Anteilsschätzern geschätzt werden können. In Abschnitt 2.3.3 werden Aussagen über die zu erwartende Präzision der zuvor eingeführten Schätzer getroffen.

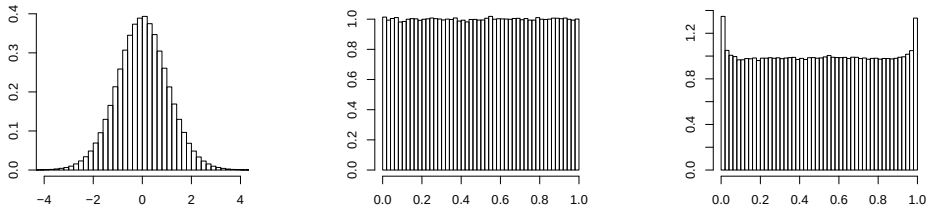
2.3.1 Signifikanzniveau

Die Grundlage der empirischen Verifikation der Fehlerrate statistischer Tests bildet der *Integraltransformationssatz* (Ewens und Grant, 2001, Kap. 1.15):

Satz 2.1 (Integraltransformationssatz). *X habe die stetige Verteilungsfunktion F. Dann ist $F(X)$ gleichverteilt auf dem Intervall* $[0, 1]$.

Für einen Beweis dieses Satzes sei auf die weiterführende Literatur, beispielsweise Büning und Trenkler (1994) oder Ewens und Grant (2001) verwiesen.

Da das Signifikanzniveau den Anteil derjenigen Tests beschreibt, deren Datenbasis nur zufällig eine Ablehnung der Nullhypothese erlaubt, kann leicht eine Methode angegeben werden, die die Verifikation einer vorgegebenen Fehlerrate α erlaubt: unter der Nullhypothese werden M_{H_0}, etwa $M_{H_0} =$ 100.000, Datensätze simuliert. Für jeden Datensatz $1 \leq k \leq M_{H_0}$ wird die Teststatistik T_k und die Wahrscheinlichkeit daß $F(T) > T_k$ gilt ermittelt (der p-Wert). Der Anteil derjenigen Replikate, deren p-Wert kleiner ist als das vorgegebene Niveau, also diejenigen Replikate, die die Bedingung $P\left(F(T) > T_k\right) < \alpha$ erfüllen, gibt das empirische Signifikanzniveau α_{emp} an.



(a) (b) (c)

Abbildung 2.4: Eine hypothesische Teststatistik sowie die Verteilung der p-Werte unter Annahme verschiedener Verteilungsfunktionen

In Formeln:

$$\alpha_{emp} = \frac{\#\{P\left(F(T) > T_k\right) < \alpha\}}{M_{H_0}}.$$

Ein Test, der das Niveau für jedes α einhält, für den also empirisches und nominales Niveau übereinstimmen, heißt *unverfälscht* (Lehmann und Romano, 2005, Kap. 3).

Ein Beipiel soll diese Methode illustrieren: Ein Test T^{20} sei t-verteilt mit 20 Freiheitsgraden. Genähert sei aber nur eine Standardnormalverteilung bekannt. Mittels einer Simulation, die den tatsächlichen zufälligen Prozess unter der Nullhypothese möglichst naturgetreu abbildet, werden M_{H_0} Replikate erzeugt – Abb. 2.4(a) zeigt exemplarisch die mögliche Verteilung der M_{H_0} Teststatistikwerte. Abbildungen 2.4(b) und 2.4(c) zeigen die Verteilungen der p-Werte unter der Annahme der beiden Verteilungsfunktionen: einmal unter der tatsächlichen t-Verteilung mit 20 Freiheitsgraden (Abb. 2.4(b)) sowie unter der Näherung der Standardnormalverteilung (Abb. 2.4(c)). Hier ist die

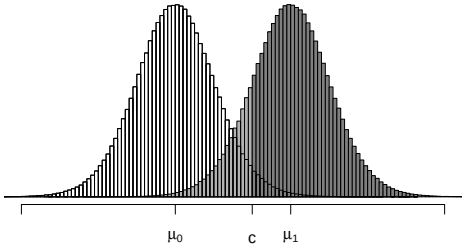


Abbildung 2.5: Verteilungen simulierter Testergebnisse unter der Nullhypothese (links) wie unter der Alternativen (rechts). Der dunkel gefärbte Bereich entspricht der empirischen Macht des Tests auf Basis der Simulation.

Verbindung zum Satz 2.1 besonders deutlich: bei Verwendung der korrekten Verteilungsfunktion sind die p-Werte gleichverteilt (Abb. 2.4(b)).

Ist dagegen die korrekte Verteilung nicht genau bekannt und dadurch die Verteilung der p-Werte wie in Abb. 2.4(c) verzerrt, so spricht man, je nach Verzerrungsrichtung, von einem *zu liberalen* oder einem *zu konservativen* Test. Konservative Tests schöpfen das Niveau nicht voll aus (nicht gezeigt), während Abb. 2.4(c) einen zu liberalen Test dokumentiert.

2.3.2 Statistische Güte

Die statistische Güte G läßt sich empirisch auf einem ähnlich direkten Wege ermitteln.

Zu den bereits vorhandenen M_{H_0} Simulationen unter der Nullhypothese werden nochmals zusätzlich M_{H_1} (etwa $M_{H_1} = 1.000$) Replikate unter einer Alternativen simuliert. Die Güte des Tests, seine *Macht*, entspricht dem Anteil derjenigen Replikate der Alternativen, deren Wert der Teststatistik größer ist als ein kritischer Wert c. Der kritische Wert wird dem $1-\alpha$ Quantil der Verteilung unter der Nullhypothese entnommen, c entspricht also dem $(1-\alpha) \cdot M_{H_0}$ten Element der sortierten Stichprobe unter H_0. Damit ist die

M	1%	10%	50%
1.000	[0.0048; 0.0183]	[0.0821; 0.1203]	[0.4686; 0.5315]
10.000	[0.0081; 0.0122]	[0.0942; 0.1061]	[0.4902; 0.5099]
100.000	[0.0094; 0.0106]	[0.0982; 0.1019]	[0.4969; 0.5031]
1.000.000	[0.0098; 0.0102]	[0.0994; 0.1006]	[0.4990; 0.5010]

Tabelle 2.1: 95%-Konfidenzintervalle für Anteilsschätzer bei steigender Fallzahl M, berechnet mit der Funktion `binom.test()` in `R` (R Development Core Team, 2005) .

empirische Güte G gegeben durch:

$$G_{emp} = \frac{\#\{T_{H_1} > c\}}{M_{H_1}}$$

Abbildung 2.5 zeigt die Histogramme der simulierten Daten, links, über μ_0 zentriert, die Replikate der Nullhypothese, rechts, die Replikate der Alternativen mit Mittelwert μ_1. Der Punkt c bezeichnet die Stelle, ab der die Nullhypothese verworfen werden kann. Der dunkel gefärbte Bereich der rechten Verteilung entspricht der empirischen Power des Testes auf Basis der Simulationen. Existieren zwei unverfälschte Tests für ein Testproblem, so ist derjenige Test der *bessere*, dessen Macht größer ist.

2.3.3 Präzision der Anteilsschätzer

Für die in den vorherigen Abschnitten vorgestellten Schätzer des empirischen Signifikanzniveaus bzw. der empirischen Güte lassen sich, unter Berücksichtigung der Anzahl der simulierten Replikate, Konfidenzintervalle angeben. Tabelle 2.1 zeigt die Konfidenzintervalle einiger ausgewählter Anteilsschätzer bei steigender Fallzahl M auf einem Konfidenzniveau von 95%. Die Werte der Tabelle sind wie folgt zu interpretieren: für ein empirisches Signifikanzniveau

von 1% liegt für $M = 100.000$ der wahre Wert mit 95-%iger Wahrscheinlichkeit im Intervall $[0.0094; 0.0106]$. Das heißt, daß das empirische Signifikanzniveau auf zwei Stellen Genauigkeit angegeben werden kann. Der wahre Wert der Güte, bei $M = 1000$ Replikaten unter der Alternativen, liegt, bei 500 signifikanten Ereignissen, zu 95% im Intervall von $[0.4686; 0.5315]$, von Genauigkeit kann hier kaum gesprochen werden. Viel mehr vermittelt der Schätzer, bei $M = 1000$ Replikaten, nur einen ungefähren Eindruck der Grössenordnung.

Kapitel 3

Gewichtung nach Marker-Informativität bei quantitativen Phänotypen – Verfahren von Haseman und Elston

Die Arbeit von Haseman und Elston (1972) ist eine der meistzitierten der genetischen Statistik und kann als eine der Wurzeln der modernen modellfreien Kopplungsanalyse für quantitative Phänotypen betrachtet werden. Aufgrund der Einfachheit, Eleganz und Robustheit des vorgestellten Ansatzes wird dieser nicht nur heute noch angewendet, sondern auch aktiv weiterentwickelt. Die Zeitschrift Human Heredity veröffentlichte eine Spezialausgabe zum drei-

ßigsten Jahrestag des Verfahrens von Haseman und Elston (Vol 55, No 2-3, 2003).

In diesem Kapitel soll der Ansatz von Haseman und Elston um Gewichtung nach Markerinformativität erweitert werden. In Abschnitt 3.1 wird eine thematisch fokussierte Einführung zur Methode von Haseman und Elston gegeben. Die mathematischen Grundlagen zur Erweiterung des zuvor beschriebenen klassischen Verfahrens durch Gewichtungselemente werden in Abschnitt 3.2 gelegt. Der eigentliche gewichtete Ansatz wird in Abschnitt 3.2.2 entwickelt. Zwei unabhängig durchgeführte Simulationsstudien zur Validierung diese Gewichtungsansatzes werden in Abschnitt 3.3 beschrieben. Aus den Ergebnissen der Simulationsstudien lassen sich dann die in Abschnitt 3.4 dargestellten Schlüsse ziehen: es besteht die Gefahr, daß Studien durch das *Design von p-Werten* über plausibel erscheinende Gewichtungsschemata manipuliert werden könnten.

3.1 Einführung

Der Ansatz von Haseman und Elston (1972) basiert zum einen auf dem Ähnlichkeitsprinzip: Geschwisterpaare die sich genotypisch ähnlich sind, sollten den zum Genort gehörigen Phänotyp entsprechend ähnlich ausprägen. Zum anderen fußt er auf der linearen Regression der phänotypische auf die genetischen Ähnlichkeit.

Für die Regression kann der Phänotypstatus sowohl als binär (krank bzw. nicht krank), als auch als quantitativ angenommen werden. In den aktuellen Arbeiten werden, wie in der Originalarbeit von 1972, quantitative Phäno-

typen verwendet. Die erste Anwendung wurde hingegen mit binären Zielgrößen (Schizophrenie) auf einer selektierten Stichprobe durchgeführt (Elston et al., 1973).

Eine formell korrekte Definition des Begriffs der phänotypischen Ähnlichkeit beinhaltet die Definition eines (Abstands-)Maßes. Das klassische Maß hierfür ist die quadrierte phänotypische Differenz, also der quadrierte euklidische Abstand der phänotypischen Ausprägungen. In der Literatur wurden andere Definitionen der phänotypischen Ähnlichkeit diskutiert (Drigalenko, 1998; Elston et al., 2000). Als Maß der genetischen Ähnlichkeit Anzahl der Allele identisch nach Herkunft (engl.: identical by descent, IBD) oder Anteil der Allele IBD (vgl. Abschnitt 2.2.1 bzw. 2.2.2) verwendet.

Ist der Steigungsparameter der linearen Regression der phänotypische auf die genotypischen Ähnlichkeit signifikant kleiner als null, so ist von Kopplung zwischen untersuchtem Marker und Phänotyp auszugehen (vgl. Ziegler, 1999, Kap. 3).

Modifikationen und Gewichtungen

Es wurden bereits viele Ansätze, die statistische Güte des Verfahrens von Haseman und Elston (1972) zu verbessern, untersucht. In dieser Arbeit sollen aber nur Erweiterungen vorgestellt werden, die auf Gewichtungsansätzen beruhen.

Versuche durch Gewichtung nach Phänotypinformation eine Verbesserung der statistschen Güte zu erreichen, wurden bereits von Amos et al. (1989) und Sham und Purcell (2001) unternommen. Ein späterer Ansatz von Jacobs

et al. (2003) beschäftigte sich auch mit den Möglichkeiten der Gewichung nach Genotypinformationen.

Gewichtung nach Phänotypinformationen

Amos et al. (1989) schlugen einen Verallgemeinerten Kleinste Quadrate Ansatz (engl.: generalized least squares, GLS) vor. Sie verwenden Gewichte, die auf der Fisher-Information der Phänotypen, gegeben den Markerinformationen, basieren. Durch die Varianzschätzung aus den vorhandenen Daten heraus ergeben sich allerdings Probleme; erst bei Fallzahlen von über 300 Geschwisterpaaren kann davon ausgegangen werden, daß die angenomme Asymptotik greift, und die Ergebnisse nicht zu liberal ausfallen. Außerdem können aufgrund der Konstruktion der Schätzer negative Varianzen auftreten (Amos et al., 1989).

Einen anderen Weg gingen Sham und Purcell (2001). Sie verwenden als abhängige Variable für die Regression von Haseman und Elston nicht die quadrierte phänotypische Differenz, sondern eine Linearkombination aus quadrierter phänotypischer Differenz und der quadrierten, mittelwertzentrierten, phänotypischen Summe. Diese Linearkombination wird gebildet indem die Phänotypen anhand der Phänotyp-Korrelation zwischen Geschwisterpaaren und Population gewichtet werden.

Beide Ansätze für die Gewichtung nach Phänotypinformationen ergaben eine deutliche Verbesserung der statistischen Güte für das Verfahren von Haseman und Elston.

Gewichtung nach IBD-Informationen

Im Gegensatz zu den im vorherigen Absatz angesprochenen Varianten wird bei der Gewichtung nach IBD-Informationen nach eventuell zusätzlich vorhandener, bisher ungenutzter, Information in den Genotypdaten gesucht.

Wie zu Beginn dieses Abschnittes dargestellt, wird in der Methode von Haseman und Elston (1972) der Anteil der Allele IBD verwendet, eine verlustbehaftete Komprimierung der zuvor aus den experimentell ermittelten Genotypdaten geschätzten IBD-Werte. Die Variabilitäten und Unsicherheiten des Anteils der Allele IBD werden dabei ignoriert (Abschnitt 2.2).

Einen Ansatz, diese Variabilität einzubeziehen, beschreiben Jacobs et al. (2003): IBD-Werte werden anhand ihres euklidischen Abstandes zur uninformativen a-priori IBD-Schätzung gewichtet. Gemessen wird der Abstand in kartesischen Koordinaten über dem de Finetti Diagramm (Abschnitt 2.2, Abb. 2.3). Zum Genetic Analysis Workshop (GAW) 13 setzten Jacobs et al. (2003) diese Art der Gewichtung nach IBD-Informationen ein. Zu diesem Zeitpunkt war allerdings noch nicht geklärt, ob Gewichtung in diesem Kontext valide ist und tatsächlich zu einer Verbesserung der statistischen Güte des Tests von Haseman und Elston (1972) führt.

Eine detailierte Herleitung der Methode, eine Analyse und Antworten auf diese offenen Fragen wird in den nachfolgenden Abschnitten gegeben.

3.2 Methoden

Das von Haseman und Elston (1972) vorgestellte lineare Modell beruht einerseits auf dem Ähnlichkeitsprinzip, einem Zusammenhang von genetischer und

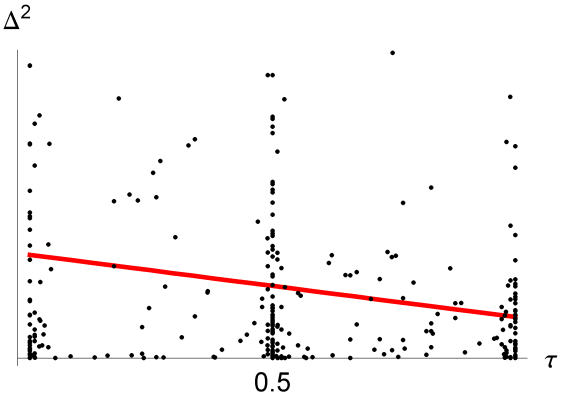


Abbildung 3.1: Illustration des Verfahrens von Haseman und Elston, 300 Geschwisterpaare Geschwisterpaare an einem quantitativen Krankheitsgenort. Informationsgehalt: $h^2 = 0.4$, $PIC = 65\%$.

phänotypischer Ähnlichkeit, andererseits auf einer gewöhnlichen linearen Regression der genotypischen Ähnlichkeit auf die phänotypische. Abbildung 3.1 zeigt exemplarisch einen solchen Zusammenhang, anhand von 300 simulierten Geschwisterpaaren. Der Informationsgehalt des Polymorphismus (engl.: Polymorphism Information Content, PIC) nach Shete et al. (2000) beträgt am betrachteten Marker etwa 65%.

In den folgenden Abschnitten wird das Ähnlichkeitsprinzip mathematisch beschrieben: zuerst für das klassische lineare Modell von Haseman und Elston, dann für das allgemeine lineare Modell (Abschnitte 3.2.1 und 3.2.2), welches die Verwendung von Gewichtungsschemata nach IBD-Information erlaubt. Die Gewichtungsschemata für das allgemeine lineare Modell werden in Abschnitt 3.2.3 eingeführt. Desweiteren erfolgt eine Betrachtung der sich ergebenden Implikationen, wenn die Regressionsrichtung umgekehrt wird, d. h. wenn man die genotypische auf die phänotypische Ähnlichkeit regressiert (Abschnitt 3.2.4). Hiermit wird ein Resultat von Schaid et al. (2003, Appendix A) verallgemeinert .

3.2.1 Lineares Modell

Zur Formulierung des linearen Modells werden zu Beginn einige Notationen eingeführt. Seien x_{1k}, x_{2k} die phänotypischen Ausprägungen des $k$ten von n unabhängigen Geschwisterpaaren, so sei $\Delta_k^2 = (x_{1k} - x_{2k})^2$ die quadrierte phänotypische Differenz des Paares. Außerdem sei $\tau_k = z_{2k} + z_{1k}/2$ der Anteil der Allele IBD des Geschwisterpaares k bei IBD-Verteilung (z_{0k}, z_{1k}, z_{2k}). Dann läßt sich mit

$$\boldsymbol{\Delta^2} = \begin{pmatrix} \Delta_1^2 \\ \vdots \\ \Delta_n^2 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & \tau_1 \\ \vdots & \vdots \\ 1 & \tau_n \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \text{ und } \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \boldsymbol{I}\right)$$

das klassische lineare Modell ohne Dominanzvarianz von Haseman und Elston wie folgt formulieren:

$$\boldsymbol{\Delta}^2 = \boldsymbol{X}\,\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad . \tag{3.1}$$

Das Modell von Haseman und Elston (3.1) ist äquivalent zur Anpassung einer Regressionsgeraden durch die Punktewolke aller Paare (τ_k, Δ_k^2), $k = 1, \ldots, n$ (vgl. Abb. 3.1). Eine notwendige Bedingung für Kopplung ergibt sich aus dem Erwartungswert von $\hat{\beta}$: es läßt sich zeigen, daß nur dann Kopplung vorliegt wenn $\hat{\beta}$ signifikant kleiner ist als 0 (Ziegler, 1999, Kap. 3). Üblicherweise wird als Test für $\beta < 0$ ein t-Test der Form

$$T_{HE} = \frac{\hat{\beta} - 0}{\sqrt{\widehat{\mathbb{V}ar}\left(\hat{\beta}\right)}} \sim t_{n-1} \tag{3.2}$$

herangezogen. Hierbei sind, wie aus der Theorie der linearen Modelle bekannt, $\hat{\gamma} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ und $\widehat{\mathbb{V}ar}(\hat{\gamma}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Der Wert der Teststatistik, T_{HE}, entstammt dann einer t-Verteilung mit $n-1$ Freiheitsgraden.

3.2.2 Allgemeines lineares Modell

Das allgemeine lineare Modell läßt die Annahme der Homoskedastie zugunsten der *Heteroskedastie* fallen. Heteroskedastie bedeutet, daß die zufälligen Fehler aus Normalverteilungen mit verschiedenen Varianzen entstammen können. Entsprechend wird die Einheitsmatrix $\boldsymbol{I}$ durch eine Diagonalmatrix $\boldsymbol{\Omega} = diag(\omega_k)$ ersetzt. Dabei können die Einträge ω_k prinzipiell beliebig, aber echt größer als 0, gewählt werden. Die Idee des Ansatzes ist es, die ω_k nicht willkürlich zuzuordnen, sondern die Informativität der verwendeten IBD-Schätzer zugrunde zu legen. Je informativer und eindeutiger eine IBD-Schätzung ist, desto kleiner sollte der zugehörige Fehler sein. Somit kann über die Heteroskedastie bezüglich der Informativität der IBD-Schätzer τ_k adjustiert werden.

Hierfür wird der Aitken-Schätzer verwendet, der die asymptotischen Eigenschaften des klassischen Schätzers $\hat{\gamma}$ auch für den Fall heteroskedastischer Varianzen erbt.

Aitken-Schätzer für das allgemeine lineare Modell

Der Unterschied zwischen dem homoskedastischen und heteroskedastischen Modell liegt in der zugelassenen Kovarianzmatrix des Fehlerterms $\boldsymbol{\varepsilon}$. Entspre-

chend wird die Matrix $\boldsymbol{\Omega}$ über eine Transformation auf die Einheitsmatrix $\boldsymbol{I}$ zurückgeführt.

Die Kovarianzmatrix des heteroskedastischen Modells $\boldsymbol{\Omega}$ hat nach Voraussetzung Diagonalgestalt bei vollem Rang, d. h. alle Einträge sind echt größer als 0. Mit der Zerlegung $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Omega}^{-\frac{1}{2}}\,\boldsymbol{\Omega}^{-\frac{1}{2}}$ läßt sich das heteroskedastische in ein homoskedastisches Modell transformieren. Dazu wird unter Annahme heteroskedastischer Fehler, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{\Omega})$, die Modellgleichung (3.1) von links mit der Matrix $\boldsymbol{\Omega}^{-\frac{1}{2}}$ multipliziert:

$$\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}^2 = \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{X}\,\boldsymbol{\gamma} + \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\varepsilon} \quad \Leftrightarrow \quad \boldsymbol{\Delta}^2_* = \boldsymbol{X}_*\,\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_* \tag{3.3}$$

Mit $(\boldsymbol{\Omega}^{-\frac{1}{2}})' = \boldsymbol{\Omega}^{-\frac{1}{2}}$ läßt sich zeigen, daß die Varianz des transformierten Modells nun der Varianz eines homoskedastischen entspricht:

$$\mathbb{V}ar\,(\boldsymbol{\varepsilon}_*) = \boldsymbol{\Omega}^{-\frac{1}{2}}\,\sigma^2\boldsymbol{\Omega}\,(\boldsymbol{\Omega}^{-\frac{1}{2}})' = \sigma^2\boldsymbol{\Omega}^{-\frac{1}{2}}(\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\Omega}^{\frac{1}{2}})\boldsymbol{\Omega}^{-\frac{1}{2}} = \sigma^2\boldsymbol{I}\,.$$

Die sogenannten *Aitken-Schätzer* für $\hat{\gamma}$ und $\widehat{\mathbb{V}ar}\,(\hat{\gamma})$ des allgemeinen linearen Modells lassen sich über die Schätzer für das lineare Modell bestimmen. Aus der Modellgleichung $\boldsymbol{\Delta}^2_* = \boldsymbol{X}_*\,\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_*$ ergibt sich mit $\boldsymbol{X}_* = \boldsymbol{\Omega}^{-\frac{1}{2}}\,\boldsymbol{X}$ und $\boldsymbol{\Delta}^2_* = \boldsymbol{\Omega}^{-\frac{1}{2}}\,\boldsymbol{\Delta}^2$:

$$\hat{\gamma} = (\boldsymbol{X}'_*\boldsymbol{X}_*)^{-1}\,\boldsymbol{X}'_*\boldsymbol{\Delta}^2_* = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\,\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Delta}^2 \tag{3.4}$$

Analog folgt für den Schätzer der Varianz von $\hat{\gamma}$:

$$\widehat{\mathbb{V}ar}\,(\hat{\gamma}) = \hat{\sigma}^2\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1} \tag{3.5}$$

Das lineare Modell ist damit ein Spezialfall des allgemeinen linearen Modells falls $\hat{\mathbf{\Omega}} \equiv \boldsymbol{I}$.

Anwendung des allgemeinen linearen Modells

Da die Matrix $\mathbf{\Omega}$ unter Berücksichtigung der genannten Voraussetzung frei wählbar ist, lassen sich Algorithmen definieren, die, unter Einbeziehung der Informativität einer IBD-Schätzung, eine Gewichtung in die Methode von Haseman und Elston einbringen. Informativen IBD-Schätzer wird eine geringere Fehlervarianz unterstellt als uninformativen IBD-Schätzern. Ziel der Gewichtung nach Markerinformation ist es, unter Einhaltung des Signifikanzniveaus, eine höhere statistische Güte als durch Verwendung des ungewichteten Ansatzes zu erreichen. Sollte dies möglich sein, wäre der nächste Schritt die Suche nach einem optimalen Gewichtungsschema, welches die maximale Güte unter allen Gewichtungsschemata, die das Signifikanzniveau einhalten, erreicht.

Wie eine Zuordnung zwischen Informativität und Gewichten, unabhänging von Optimalitätsbedingungen aussehen könnte, wird im nächsten Abschnitt gezeigt.

3.2.3 Gewichtungsschemata für das allgemeine lineare Modell

In Abschnitt 2.2.4 wurden bereits Maße für die Informativität von IBD-Verteilungen vorgestellt. In diesem Abschnitt werden diese Informationsma-

ße dazu verwendet, verschiedene Schätzer $\hat{\mathbf{\Omega}}$ für die Kovarianzmatrix $\mathbf{\Omega}$ zu konstruieren.

Auch wenn die Matrix $\hat{\mathbf{\Omega}}$ und damit das Informativitätsmaß unter den im vorherigen Abschnitt genannten Einschränkungen frei wählbar ist: es existiert nur eine wahre Varianzmatrix, alles andere sind Näherungen. Wie gut diese Näherungen sind und ob sich damit überhaupt zulässige Ergebnisse ergeben, soll mit den bereits in Abschnitt 2.3 erarbeiteten Methoden im weiteren Verlauf dieses Kapitels überprüft werden. Nach einer Beschreibung, wie sich Gewichtungsschemata aus den in Abschnitt 2.2.4 gezeigten Informationsmaßen ableiten lassen, folgen einige Schemata im Detail.

Vom Informationsmaß zum Gewichtungsschema

Sei $\mathbf{\Omega}$ eine Diagonalmatrix mit $\mathbf{\Omega} = diag(\omega_k)$ mit $w_k > 0$ für alle k, dann ist $\mathbf{\Omega}^{-1} = diag(\omega_k^{-1})$, $k = 1, \ldots, n$ die Inverse von $\mathbf{\Omega}$. Die Inverse der Kovarianzmatrix $\mathbf{\Omega}^{-1}$ ist sowohl im Schätzer für die Steigung der Regressionsgeraden des allgemeinen linearen Modells (3.4) enthalten, als auch im Schätzer für deren Varianz (3.5). Somit ist ausreichend, für jedes der n Geschwisterpaare, jeweils einen Fehlervarianzschätzer $\hat{\omega}_k$, $k = 1 \ldots n$, anzugeben derart, daß die Fehlervarianz für informative Marker geringer sein soll, als für uninformative. Das heißt: sei d_k die Informativität des $k$ten Geschwisterpaares, dann wird $\hat{\omega}_k = d_k^{-1}$ als ihr Gewichtungsfaktor verwendet.

Gewichtung nach euklidischem Abstand

Der Abstand einer IBD-Schätzung v von der uninformativen a-priori Schätzung u für Geschwisterpaare ist in der Ebene des de Finetti-Diagrams durch

Gleichung (2.2) aus Abschnitt 2.1 gegeben. Entsprechend ist, nach obiger Vorschrift $\hat{\omega}_{euklid}$ definiert als

$$\hat{\omega}_{euklid}\,(v_0, v_1, v_2) = \left(\frac{(v_0 - v_2)^2}{3} + \frac{(1 - 2v_1)^2}{4} \right)^{-\frac{1}{2}} \quad . \tag{3.6}$$

Gewichtung nach Entropie

Je unähnlicher zwei IBD-Schätzungen sich sind, desto größer ist ihre relative Entropie. Die relative Entropie zwischen einer IBD-Schätzung v und der uninformativen a-priori IBD-Schätzung u ist in Abschnitt 2.2, Gleichung (2.4), gegeben. Entsprechend sind $\omega_{shannon}$ und $\omega_{modshannon}$ gegeben durch

$$w_{shannon}(v) = \left(\frac{v_0 \log(4\,v_0) + v_1 \log(2\,v_1) + v_2 \log(4\,v_2)}{\log 4} \right)^{-1} \tag{3.7}$$

beziehungsweise

$$w_{modshannon}(v) = \left(\frac{v_0 \log(3\,v_0) + v_1 \log(3\,v_1) + v_2 \log(3\,v_2)}{\log 3} \right)^{-1} \quad . \tag{3.8}$$

3.2.4 Invertiertes lineares Modell

Im klassischen linearen Modelle wird davon ausgegangen, daß die Regressoren fest vorgegeben sind. Die Matrix $\boldsymbol{X}$ der Regressoren folgt damit einem Studiendesign und wird daher auch *Designmatrix* genannt. Die zugehörigen Regressanden $\boldsymbol{y}$ werden erst gemessen, nachdem die Regressoren für das Modell festgelegt wurden.

Folgt man dieser Argumentation im Falle des Ansatzes von Haseman und Elston, in dem zuerst der Regressand (die quadrierte Phänotypische

Differenz, Δ^2) beobachtet und dann der Regressor (der Anteil der Allele IBD, τ) gemessen wird, so müßte man eigentlich von einem Regressionsmodell der Form $\tau_k = \alpha + \beta\Delta_k^2 + \varepsilon_k$, statt des Modells wie in (3.1) dargestellt, ausgehen.

Schaid et al. (2003) zeigten, daß für die einfache lineare Regression und einem Test auf den Steigungsparameter β beide Sichtweisen äquivalent sind. hier werden die Resultate von Schaid et al. (2003) verallgemeinert.

Äquivalenz zwischen Regression und Umkehrregression

Schaid et al. (2003) zeigten die Äquivalenz zwischen Regression und Umkehrregression nicht anhand des Tests wie in Gleichung (3.2) dargestellt, sondern anhand des, auf dem Korrelationskoeffizienten r beruhenden, äquivalenten, Unabhängigkeitstests (Collins und Morton, 1995):

$$t_{n-2} = \frac{\sqrt{n-2}\, r_{xy}}{\sqrt{1 - r_{xy}^2}} \tag{3.9}$$

Mit dem allgemeinen linearen Modell (3.3), einer beliebigen Kovarianzmatrix $\boldsymbol{\Omega}$ und den Definitionen von $\boldsymbol{X}_* = \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{X}$ und $\boldsymbol{\Delta}_*^2 = \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Delta}^2$ sei $x_k^* := \frac{\tau_k}{\sqrt{\omega_k}}$ sowie $y_k^* := \frac{\Delta_k^2}{\sqrt{\omega_k}}$. Dann folgt analog zu Schaid et al. (2003) mit den bekannten Darstellungen für Varianz

$$\widehat{\mathbb{V}ar}\,(x^*) = \frac{1}{n}\sum_{i=1}^{n}(x^* - \bar{x}^*)^2 \quad \text{bzw.} \quad \widehat{\mathbb{V}ar}\,(y^*) = \frac{1}{n}\sum_{i=1}^{n}(y^* - \bar{y}^*)^2$$

und Kovarianz

$$\widehat{\mathbb{C}ov}\,(x^*, y^*) = \frac{1}{n}\sum_{i=1}^{n}(x^* - \bar{x}^*)\,(y^* - \bar{y}^*)$$

der Korrelationskoeffizient

$$\hat{r}_{x^*y^*} = \frac{\widehat{\mathbb{C}ov}\,(x^*y^*)}{\sqrt{\widehat{\mathbb{V}ar}\,(x^*)\,\widehat{\mathbb{V}ar}\,(y^*)}} \quad .$$

Mit diesen Angaben ergibt sich der Unabhängigkeitstest zwischen x^* und y^* für das allgemeine lineare Modell zu

$$t_{n-2} = \frac{\sqrt{n-2}\,\hat{r}_{x^*y^*}}{\sqrt{1-\hat{r}^2_{x^*y^*}}} \quad .$$

Offensichtlich sind sowohl der Korrelationskoeffizient als auch die Teststatistik symmetrisch in x^* und y^*. Damit ist das Testergebnis invariant zur Regressionsrichtung. Setzt man ω_k in der Definition von x^* bzw. y^* zu 1, und somit implizit $\boldsymbol{\Omega} = \boldsymbol{I}$, so erhält man exakt das von Schaid et al. (2003) vorgestellte Ergebnis.

3.3 Validitätsprüfung

In der Einführung zu diesem Kapitel wurde die Frage nach der Validität von Gewichtung nach Marker-Informativität im Ansatz von Haseman und Elston aufgeworfen. Zwei unabhängige Simulationsstudien sollen hierüber Auskunft geben: sowohl eigenhändig simulierte Daten als auch unabhängig erzeugte Datensätze wurden untersucht.

3.3.1 Simulationsstudien I

Zur Validierung des Signifikanzniveaus der mit dem allgemeinen linearen Modell eingesetzten Gewichtungsschemata sowie zum Vergleich der statistischen Güte zwischen klassischem und allgemeinem linearen Modell wurden Simulationsstudien durchgeführt.

Phänotyp und Population

Als Simulationsgrundlage dienten Kernfamilien bestehend aus einem Elternpaar und zwei gemeinsamen Nachkommen. Der quantitative Phänotyp der $i$ten Person der $k$ten Familie x_{ik} wurde zur Simulation in mehrere additive Komponenten aufgeteilt (vgl. Falconer und Mackay, 1996). Dazu gehören der Populationsmittelwert μ, ein Hauptgeneffekt g_{ik}, ein von den Individuen unabhängiger, aber in der $k$ten Familie präsenter gemeinsamer Umwelteffekt E_k sowie ein personenbezogener Fehlerterm ε_{ik}:

$$x_{ik} = \mu + g_{ik} + E_k + \varepsilon_{ik} \quad .$$

Der Populationsmittelwert μ wurde ohne Beschränkung der Allgemeinheit als verschwindend angenommen. Der Hauptgeneffekt g_{ik} wurde sowohl durch den Genotyp des bi-allelischen Krankheitsgenorts, als auch über das zuvor festgelegte Vererbungsmodell (additiv, dominant oder rezessiv) bestimmt. Zur Gesamtvarianz des Phänotyps trug er 20% bei. Der gemeinsame Familieneffekt E_k, ein zufälliger, normalverteilter Effekt, der jedem Familienmitglied gleichermaßen zugewiesen wurde, erklärt weitere 30%. Die restlichen 50% der Varianz wurden durch zufällige Effekte bestimmt. Der Zufallseffekt ε_{ik}

	m01	m02	m03	m04	m05	m06	m07
cM	0	7.1	13.6	20.9	28.4	36.7	43.3
#	12	7	12	9	11	8	8
	m08	m09	m10	m11	m12	m13	
cM	51.2	57.9	64.7	71.1	78.8	84.7	
#	11	6	8	8	11	7	

Tabelle 3.1: Markerverteilung für Multi-Marker Simulation.

entspricht hierbei einer normalverteilten Zufallsvariablen. Es wurden sowohl Zwei- als auch Mehrpunktmodelle in Betracht gezogen.

Das Zweipunktmodell wurde mit einem biallelischen Krankheits- und einem Markergenort mit zehn gleichhäufigen Allelen durchgeführt. Der PIC-Wert des Markers beträgt in diesem Fall jeweils etwa 90% (Shete et al., 2000). Der Krankheitsgenort wurde zur Fehlersimulation ungekoppelt, zur Bestimmung der Güte unter voller Kopplung, simuliert. Für den ungekoppelten Fall wurden 100.000 Replikationen mit jeweils 300 Kernfamilien erstellt, im gekoppelten Fall 1.000 Replikationen, ebenfalls mit je 300 Kernfamilien.

Um ein realen Anwendungsfall zu simulieren wurde auch eine Simulation mit 13 verschieden informativen Markern generiert (Tab. 3.1). Die Abstände zwischen den Markern als auch die Anzahl N der Allele wurden zufällig ausgewählt. Die Wahrscheinlichkeit p_{ij} des $j$ten Allels am $i$ten Marker ist definiert durch $p_{ij} = 2^{N_i - j}/(2^{N_i} - 1)$. Damit sank der Informationsgehalt in Het und PIC am untersuchten Marker auf etwa 40% bzw. 66%. Zur Fehlersimulation wurde der Krankheitslocus einmal ungekoppelt, zur Bestimmung der Güte vollständig mit Marker m07 gekoppelt, simuliert. Die Anzahl der Replikate unter der Nullhypothese betrug hier 75.000, unter der Alternativen 1.000. Für jedes Replikat wurden auch hier 300 Kernfamilien erzeugt.

Software

Sämtliche Monte-Carlo Simulationen wurden mit `SIBSIM` durchgeführt (Franke et al., 2006). Zur Analyse der simulierten Daten wurde ein Entwicklerzweig des Softwarepakets S.A.G.E. (2004) verwendet. Die Entwickler von S.A.G.E. implementierten die in dieser Arbeit vorgestellten Methoden in `SIBPAL`, alle hier vorgestellten Resultate wurden mit dieser Software errechnet. Zur Kontrolle und Verifikation der Ergebnisse wurde ein Teil der Methoden unabhängig von S.A.G.E. ein zweites Mal in `R` (R Development Core Team, 2005) implementiert.

Sämtliche hier genannte Software wurde in einer 64 CPU-Domain des SUN Fire 15000 Servers des Universitätsklinikums Schleswig-Holstein unter Solaris 9 eingesetzt und verwendet.

Empirische p-Werte

Asymptotische p-Werte wurden über die t-Verteilung mit $n-1$ Freiheitsgraden bestimmt, empirische p-Werte über ein Resamplingverfahren (10.000 Wiederholungen). Die zu einem Signifikanzniveau α gehörende statistische Güte wurde über das obere α-Fraktil der asymptotischen bzw. der empirischen Verteilung der p-Werte unter der Nullhypothese bestimmt.

Ergebnisse

Für das Zweipunktmodell konnten die Auftretenswahrscheinlichkeiten der insgesamt nur sieben möglichen IBD-Verteilungen für den ungekoppelten Fall ermittelt werden. Tabelle 3.2 faßt die Verteilungen und ihre jeweiligen Wahr-

Tabelle 3.2: IBD-Verteilungen und ihre Auftretenswahrscheinlichkeiten in Zweipunkt-Analysen. Bei Zweipunkt-Analysen mit erkrankten Geschwisterpaaren können nur sieben verschiedene IBD-Verteilungen auftreten, die Tabelle gibt die Wahrscheinlichkeit an, mit der ein Marker mit r gleichhäufigen Allelen die jeweiligen IBD-Verteilungen induziert.

IBD Wert	Wahrscheinlichkeit
$(1,0,0)$	$\frac{r^3-2r^2+1}{4r^3}$
$(0,1,0)$	$\frac{(r-1)^2}{2r^2}$
$(0,0,1)$	$\frac{r^3-2r^2+1}{4r^3}$
$(\frac{1}{2},\frac{1}{2},0)$	$\frac{r-1}{r^2}$
$(\frac{1}{2},0,\frac{1}{2})$	$\frac{r-1}{2r^3}$
$(0,\frac{1}{2},\frac{1}{2})$	$\frac{r-1}{r^2}$
$(\frac{1}{2},\frac{1}{4},\frac{1}{2})$	$\frac{1}{r^2}$

scheinlichkeiten zusammen. Aufgrund der limitierten Möglichkeiten bei der IBD-Verteilung, insbesondere bei angestrebten 10 Allelen am Markergenort und der daraus resultierenden hohen Informativität wurde auf eine weitere Auswertung der Simulation des Zweipunktmodells verzichtet.

Die Ergebnisse der Mehrpunktsimulationen sind, nach Fehlerhäufigkeiten 1. Art und statistischer Güte getrennt, in den Tabellen 3.3 und 3.4 zusammengestellt. Abbildung 3.2 zeigt eine Zusammenfassung in graphischer Form.

Tabelle 3.3 zeigt, daß der klassische, ungewichtete Ansatz von Haseman und Elston bei asymptotischer Bestimmung der p-Werte das nominelle Niveau sehr gut einhält. Im Gegensatz hierzu sind sowohl die Erweiterung durch Gewichtung nach euklidischem Abstand als auch nach der Shannon'schen Information deutlich zu liberal (6.1% bzw. 8.9% Fehlerhäufigkeit bei einem vorgegebenen Signifikanzniveau von $\alpha = 5\%$). Werden die p-Werte dagegen empirisch über Resampling ermittelt, halten alle Methoden das vorgegebene Niveau. Abbildung 3.2 zeigt diesen Zusammenhang als Verzerrung der ursprünglichen Gleichverteilung der p-Werte: die Teilabbildungen (a), (c)

	Empirisches α		
Nominelles α	ungewichtet	w_{euklid}	$w_{shannon}$
Asymptotische p-Werte			
0.0001	0.00011	0.00029	0.00093
0.001	0.00079	0.00152	0.00499
0.01	0.00976	0.01397	0.02691
0.05	0.04947	0.06131	0.08909
Empirische p-Werte			
0.0001	0.00000	0.00000	0.00000
0.001	0.00099	0.00100	0.00097
0.01	0.00977	0.00972	0.00995
0.05	0.04896	0.04891	0.04865

Tabelle 3.3: Fehlerhäufigkeiten von klassischen und gewichteten Ansätzen in Mehrpunktsimulationen unter einem additiven Modell. Es wurden je 300 Kernfamilien in 75.000 Replikaten unter der Nullhypothese betrachtet.

	Empirische Güte		
Nominelles α	ungewichtet	w_{euklid}	$w_{shannon}$
Asymptotische p-Werte			
0.0001	0.008	0.011	0.036
0.001	0.035	0.056	0.109
0.01	0.166	0.197	0.260
0.05	0.378	0.420	0.480
Empirisch p-Werte			
0.0001	0.009	0.009	0.006
0.001	0.047	0.047	0.039
0.01	0.172	0.167	0.166
0.05	0.378	0.378	0.367

Tabelle 3.4: Statistische Güte klassischer und gewichteter Ansätzen in Mehrpunktsimulationen unter einem additiven Modell. Es wurden je 300 Kernfamilien in 75.000 Replikationen unter H_0 und in 1.000 Replikaten unter der H_1 betrachtet.

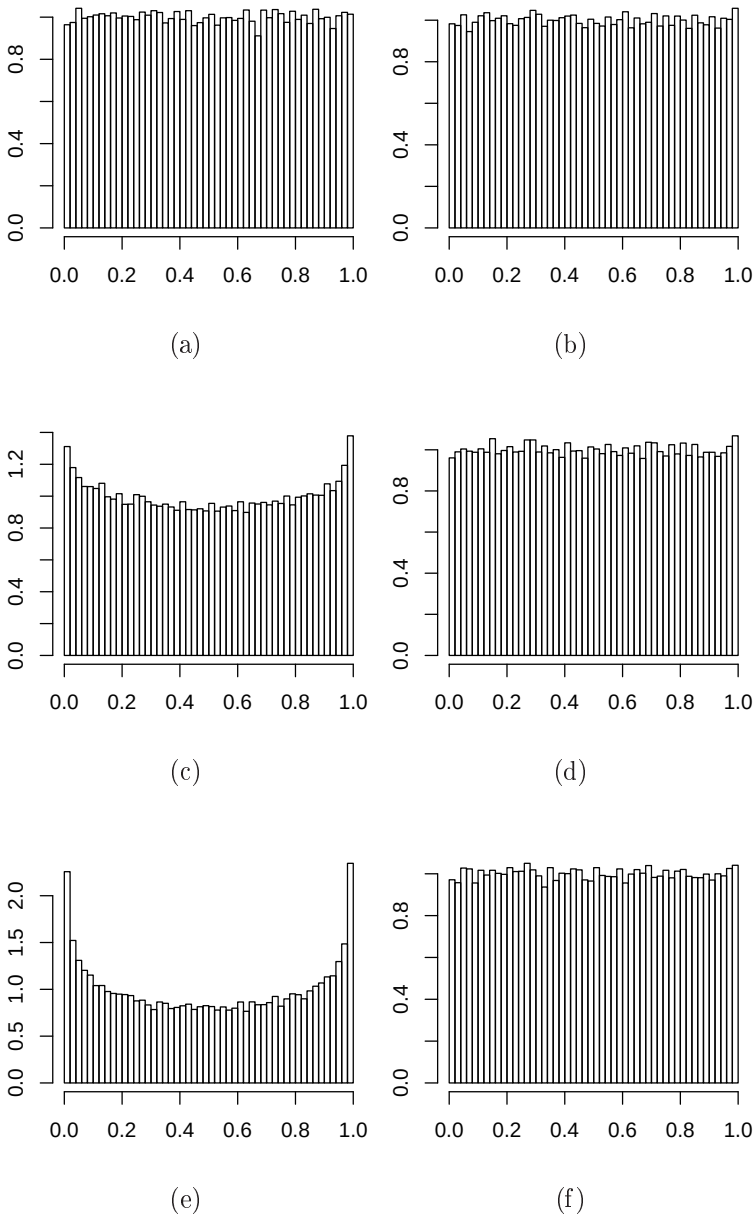


Abbildung 3.2: Verteilungen der p-Werte bei Verwendung des Ansatzes von Haseman und Elston, ungewichtet (a,b), gewichtet nach euklidischem Abstand (c,d) und gewichtet nach der Shannon'schen Information (e,f). Die Abbildungen (a,c,e) zeigen die Verteilungen der asymptotischen p-Werte, die Abbildungen (b,d,f) die entsprechend empirisch ermittelten p-Werte (vergl. Tab. 3.3).

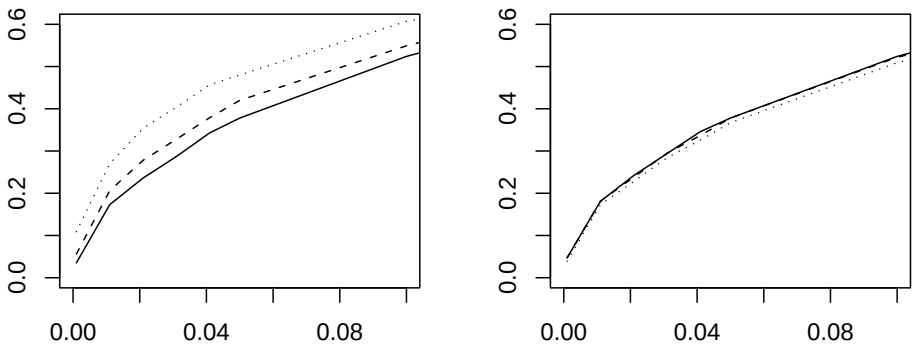


Abbildung 3.3: Statistische Güte des Ansatzes von Haseman und Elston für $\alpha < 10\%$; klassischer Ansatz (durchgezogene), gewichtet nach euklidischem Abstand (gestrichelt) und gewichtet nach der Shannon'schen Information (gepunktet). Die linke Graphik wurde unter Verwendung der der asymptotischen Verteilungen erstellt, die rechte unter Verwendung der empirischen (vergl. Tab. 3.4.)

und (e) zeigen die Verteilungen der asymptotischen p-Werte, ihre Pendants (b), (d) und (f) die entsprechenden Verteilungen der empirischen p-Werte (vgl. Abschnitt 2.3.1, Abbildung 2.4). Obwohl die Fehlerwahrscheinlichkeiten durch Gewichtung deutlich inflationiert ist, wurde auch ein Vergleich der Güte vorgenommen. Tabelle 3.4 zeigt, daß sich unter Verwendung von asymptotischen p-Werten die Wahrscheinlichkeit Kopplung, zu entdecken, durch Gewichtung nur moderat erhöht (38% zu 48% zwischen ungewichtetem Ansatz und Gewichtung nach Shannon'schem Informationsgehalt bei einem Signifikanznievau von nominal 5%), man sich diese zusätzliche Güte aber mit einer fast verdoppelten Wahrscheinlichkeit eines Fehlers 1. Art erkauft. Bei Verwendung der das Niveau einhaltenden empirischen p-Werte, lassen sich keine Veränderungen der statistischen Güte der Ansätze beobachten, sie liegt bei einem Signifikanzniveau von 5% in jedem der betrachteten Fälle bei etwa 38% (vgl. Tab. 3.4 sowie Abb. 3.3).

Die Ergebnisse bleiben Veränderungen gegenüber der Anzahl der simulierten Kernfamilien pro Replikat bzw. gegenüber einer Erhöhung des genetischen Effektes invariant (nicht gezeigt).

3.3.2 Simulationsstudien II

Als zweiter Datensatz wurden die simulierten Daten des Genetic Analysis Workshop (GAW) 14 herangezogen. Ziel war auch hier, die asymptotischen Verteilungen der klassischen sowie verschiedener gewichteter Methoden zu überprüfen. Exemplarisch für die hier hergeleiteten Gewichtungsschemata wurde die Gewichtung nach euklidischem Abstand betrachtet. zusätzlich wurde zum Vergleich die bereits zum GAW 13 eingeführte Variante von Jacobs et al. (2003) verwendet:

$$D\left(f_0, f_1, f_2\right) = \begin{cases} \frac{1}{3}\left(1 - \frac{2f_0}{1-f_1}\right)^2 + \frac{1}{4}\left(1-2f_1\right)^2 & \text{, falls } f_1 < 1 \\ \frac{1}{4} & \text{, sonst} \end{cases} \tag{3.10}$$

Phänotyp und Population

Die genaue Zusammensetzung des fiktiven Phänotyps der *Kofendrerd Personality Disorder* (KPD) wurde in der Beschreibung der GAW-Daten angegeben (Bailey-Wilson et al., 2005). Seine exakte Repräsentation ist hier allerdings nicht von Bedeutung. Als Datengrundlage wurden 100 Replikate des Chromosoms 4 der imaginären Bevölkerung von Aipotu ausgewählt. Die Auswahl erfolgte nach Entblindung anhand der Phänotypbeschreibung derart, daß keinerlei positive Ergebnisse auf dem gewählten Chromosom zu erwarten waren. Sämtliche Kopplungsbefunde mußten also rein zufällig, ohne

genetischen Hinergrund, entstanden sein. Der Datensatz aus Aipotu bestand aus 100 erweiterten Kernfamilien, einem Elternpaar mit jeweils mehreren gemeinsamen Nachkommen, von diesen Nachkommen waren mindestens zwei erkrankt.

Methoden

Um die asymptotischen Verteilungen der Verfahren nach Haseman und Elston mit Gewichtung auf Verzerrungen zu überprüfen, wurde darauf getestet, ob unter der Nullhypothese der Anteil der signifikanten Ergebnisse signifikant erhöht war. Das Heißt, daß für jede genetische Position für jedes Replikat jeder Test berechnet und die Häufigkeit der signifikanten Ergebnisse bestimmt wurde. Hierbei wurde grundsätzlich ein Signifikanzniveau von 5% zugrunde gelegt. Der Test der Validität bestand darin festzustellen, ob signifikante Abweichungen zwischen den ermittelten Häufigkeiten und der zu erwartenden Anzahl von (hier) 5 falsch-positiven Ergebnissen bestehen.

Um die Korrelation zwischen Markern abbilden zu können, wurden Generalisierte Schätzgleichungen (engl.: Generalised Estimating Equations, GEE) mit Autoregressionsstruktur (AR(1)) verwendet (vgl. Ziegler et al., 1998). Für jeden Ansatz wurden hiermit ein Punktschätzer sowie ein Konfidenzintervall für den Anteil der siginifikanten Ergebnisse berechnet.

Desweiteren soll gezeigt werden, daß die absolute Anzahl der signifikanten Ergebnisse bei den gewichteten Verfahren höher liegt als beim klassischen Ansatz. Hierzu wurde die Anzahl der genetischen Positionen, die einen p-Wert kleiner $\alpha = 5\%$ aufweisen über das ganze Chromosom aufsummiert, dann über alle 100 Replikate ein Wilcoxon-Vorzeichen-Rang-Test berechnet.

Sofern die gewichteten Varianten zu liberal sind, sollte auch die absolute Anzahl der signifikanten Ergebnisse höher sein als beim klassischen Ansatz.

Verwendete Software

Zur Analyse der simulierten Daten wurde derselbe Entwicklerzweig des Softwarepakets S.A.G.E. (2004) verwendet wie zuvor, insbesondere die Programme `GENIBD` und `SIBPAL` kamen zum Einsatz. Die GEEs wurden mit `SAS` in der Prozedur `genmod` berechnet.

Ergebnisse

Tabelle 3.5 zeigt, daß unter nicht-uniformer Gewichtung die relative Anzahl der signifikanten Ergebnisse mindestens so hoch oder höher liegt, als unter uniformer Gewichtung. Diese subjektive Einschätzung wird durch die GEE bestätigt: Der Anteil $\hat{p}$ signifikanter Ergebnisse übersteigt das nominale Niveau von $\alpha = 5\%$ deutlich. Für die Gewichtung nach euklidischem Abstand wird $\hat{p} = 0.0685$ mit einem 95% Konfidenzintervall von $0.0546 - 0.0823$ angegeben, für die Gewichtung nach Jacobs et al. (2003) ein Punktschätzer von $\hat{p} = 0.0634$ und ein Konfidenzintervall von $0.0562 - 0.0706$. Im Gegensatz hierzu sind keine Abweichungen für den klassischen Fall festzustellen: $\hat{p} = 0.0552$ mit einem Konfidenzintervall von $0.0431 - 0.0673$. Desweiteren ist die absolute Anzahl der signifikanten Ergebnisse unter Gewichtung gegenüber dem klassischen, ungewichteten Fall signifikant erhöht, $p = 3.3 \cdot 10^{-7}$ für die euklidische Gewichtung, $p = 7.1 \cdot 10^{-6}$ für die Gewichte nach Jacobs et al. (2003).

Locus	D04S1528	D04S1529	D04S1530	D04S1531	D04S1532	D04S1533	D04S1534	D04S1535	D04S1536	D04S1537	D04S1538	D04S1539	D04S1540	D04S1541	D04S1542	D04S1543	D04S1544	D04S1545	D04S1546	D04S1547	D04S1548	D04S1549
a)	8	5	5	11	9	5	6	4	7	4	3	1	2	3	6	4	3	6	6	7	6	7
b)	9	6	6	12	11	7	8	4	8	5	3	1	2	4	6	5	5	10	10	6	8	9
c)	9	5	6	11	11	6	7	4	8	4	3	1	2	4	6	5	3	8	9	6	8	9

Locus	D04S1550	D04S1551	D04S1552	D04S1553	D04S1554	D04S1555	D04S1556	D04S1557	D04S1558	D04S1559	D04S1560	D04S1561	D04S1562	D04S1563	D04S1564	D04S1565	D04S1566	D04S1567	D04S1568	D04S1569	D04S1570	D04S1571
a)	7	11	9	5	6	4	7	4	3	1	2	3	6	4	3	6	6	6	3	6	3	4
b)	10	12	11	7	8	4	8	5	3	1	2	4	6	5	5	10	10	6	4	6	4	7
c)	9	11	11	6	7	4	8	4	3	1	2	4	6	5	3	8	9	6	4	6	3	6

Tabelle 3.5: Absolute Häufigkeiten einseitiger p-Werte kleiner $\alpha = 5\%$ in 100 Replikaten des Chromosoms 4 der Bevölkerung von Aipotu.

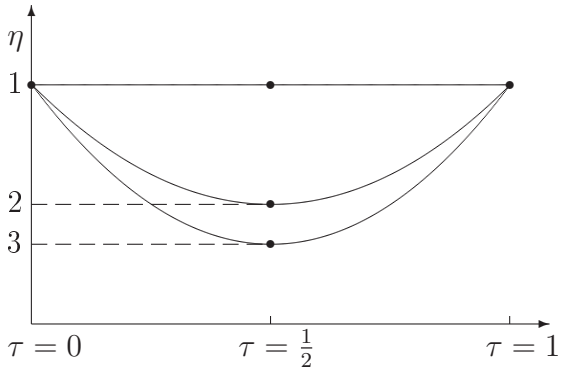


Abbildung 3.4: Parametrisierte Gewichtungsfunktion bei verschiedenen Werten des Parameters η: $\eta = 1$ (ungewichtet, klassischer Ansatz), $\eta = 2$ und $\eta = 3$.

Die hier dargestellten Ergebnisse sind invariant zur Fallzahl, d. h. paarweise Vereinigung der Replikate verändert die Ergebnisse nicht.

3.4 Design von p-Werten

Auf der Basis der selben Prinzipien läßt sich ein Verfahren konstruieren, wie p-Werte für das Verfahren von Haseman und Elston mittels Gewichtung modelliert werden können. Hierzu wird, wie zuvor, eine neue, plausibel erscheinende, Gewichtungsfunktion definiert. Diese neue, bisher nicht gezeigte, Gewichtungsfunktion enthält einen zusätzlichen Skalierungsparameter mit dessen Hilfe sich die asymptotische Testentscheidung steuern läßt.

3.4.1 Parametrisierte Gewichtungsfunktion

Unter Berücksichtigung der Mehrdeutigkeit des Anteils der Allele IBD, die ein Geschwisterpaar gemeinsam haben kann (vgl. Abschnitt 2.2, Abbildun-

gen 2.1 und 2.2), läßt sich analog zu bisherigen Argumentationen feststellen, daß ein eindeutiger Anteil Allele IBD von $\tau = 0$ oder $\tau = 1$ ein höheres Gewicht erhalten sollte als beispielsweise ein Anteil Allele IBD von $\tau = \frac{1}{2}$. Das Verhältnis der Informativität der eindeutigen gegenüber den mehrdeutigen Anteilsschätzern, also wieviel mehr Information $\tau = 0$ oder $\tau = 1$ im Vergleich zu $\tau = \frac{1}{2}$ einbringt, sei mit einem Skalierungsparameter η bezeichnet (vgl. Abb. 3.4). Je größer das Verhältnis $\eta \geq 1$ zwischen informativen und uninformativen IBD-Anteilen veranschlagt wird, desto geringer wird das Gewicht mit dem uninformative Geschwisterpaare eingehen. Die Gewichte der zwischen den Nebenbedingungen liegenden IBD-Anteile werden über eine interpolierte Parabel bestimmt (Abb. 3.4).

Formal zusammengefaßt lauten die Nebenbedingungen für die Gewichtungsfunktion $w(\tau, \eta)$:

$$w(0, \eta) = 1 \qquad w(\frac{1}{2}, \eta) = \frac{1}{\eta} \qquad w(0, \eta) = 1,$$

wobei über eine Parabelanpassung die vollständige Funktion bestimmt wird:

$$w(\tau, \eta) = 4 \cdot \left(1 - \frac{1}{\eta}\right) \cdot \left(\tau^2 - \tau\right) + 1 \quad . \tag{3.11}$$

Wird der Parameter $\eta = 1$ gewählt, ergibt sich das klassische Verfahren nach Haseman und Elston, $0 < \eta < 1$ entspräche einer Umkehrung der Argumentation, d. h. $\tau = \frac{1}{2}$ würde höher bewertet (als informativer angesehen) als $\tau = 0$ oder $\tau = 1$.

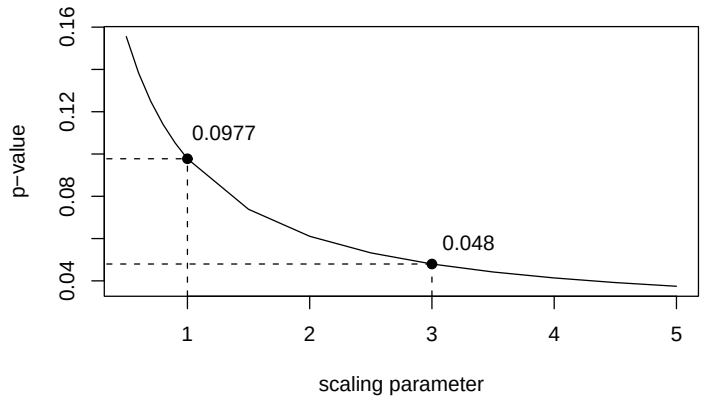


Abbildung 3.5: Entwicklung des asymptotischen p-Werts unter Verwendung einer parametrisierten Gewichtungsfunktion. Hier dargestellt ist der Marker D04S0165 von Chromosom 4 des dritten Replikats des simulierten Datensatzes aus Aipotu (Bailey-Wilson et al., 2005; Franke et al., 2005).

3.4.2 Ergebnisse

Wird das parametrisierte Gewichtungsschema (3.11) im allgemeinen linearen Modell eingesetzt (Gleichungen (3.2) bis (3.5)), so ist offensichtlich, daß T_{HE} nicht nur von den Daten und den Gewichten, sondern auch direkt von η abhängt – die Teststatistik ist damit eine Funktion des Skalierungsparameters η. Wird dieser freie Parameter im richtigen Maße angepasst, so läßt sich, unter Annahme der t-Verteilungsasymptotik, jedes gewünschte Ergebnis erzielen. Abbildung 3.5 zeigt am Beispiel eines beliebig ausgewählten Genorts der simulierten Daten des GAW 14 (vgl. Abschnitt 3.3.2) die Funktion des p-Wertes in Abhängigkeit des Skalierungsparameters η. Für ein angenommenes Signifikanzniveau von $\alpha = 5\%$ würde die Festlegung, daß die eindeutigen Anteilsschätzer ($\tau = 0$ bzw. $\tau = 1$) etwa dreimal mehr Informationen tragen

als die mehrdeutigen ($\tau = \frac{1}{2}$), ausreichen, um ein asymptotisch-signifikantes Ergebnis zu erzielen ($p = 0.048$ gegen $p = 0.0977$ im ungewichteten Fall).

Kapitel 4

Gewichtung nach Marker-Informativität bei binären Phänotypen – Mittelwertstests für erkrankte Geschwisterpaare

Nachdem im vorherigen Kapitel die Gewichtung nach Markerinformativität unter Verwendung von quantitativen Phänotypen untersucht wurde, soll in diesem Kapitel das Augenmerk auf Ansätze für binäre Phänotypen gelegt werden. Schork und Greenwood (2004) berichten, daß für eine Analyse erkrankter Geschwisterpaare eine Gewichtung nach Markerinformativität verwendet werden sollte. Ein Ansatz, dies umzusetzen, wird nachfolgend untersucht.

Abschnitt 4.1 dieses Kapitels beschreibt einführend die klassischen Mittelwerttests sowie einige ihrer, auf Gewichtungsformen basierende, Erwei-

terungen. Eine mathematische Beschreibung der Mittelwerttests als Score- und Waldtests erfolgt in Abschnitt 4.2. Es werden sowohl die klassischen Testformen als auch ein neuer, nach Markerinformativität gewichteter, Ansatz hergeleitet. Zur Überprüfung der Validität wurden Simulationsstudien durchgeführt. Diese werden in Abschnitt 4.3 vorgestellt. Abschnitt 4.4 geht auf die Anwendung des neuen Ansatzes auf reale Datensätze ein: es wurden die Daten von Risch (1990), und der Datensatz von Mein et al. (1998) reanalyisert.

4.1 Einführung

Wie in der Methode von Haseman und Elston (1972) beruhen die erkrankten Geschwisterpaar-Verfahren (engl.: affected sib pair, ASP) auf der Annahme, daß sich erkrankte Geschwister nicht nur phänotypisch ähneln, sondern auch an dem die Krankheit bedingenden Genort genotypisch ähnlich sein sollten.

Betrachtet wird bei diesen Verfahren grundsätzlich nur der Krankheitsstatus, also ob eine Person als erkrankt oder nicht erkrankt eingestuft wird. Eine quantitative Bewertung geht nicht ein. Als Maß für die genotypische Ähnlichkeit wird wie im vorherigen Kapitel der Anteil der Allele identisch nach Herkunft (engl.: identical by descent, IBD) verwendet (vgl. Abschnitt 2.2.2).

Die nachfolgend dargestellten Mittelwerttests könnten, wie Ziegler und König (2006, Kap. 7) ausführlich diskutieren, als Likelihood-Quotienten-Tests eingeführt werden. Für diese Arbeit wurde der, in der Praxis üblichere, Zugang über Wald- bzw. Score-Tests gewählt.

4.1.1 Klassische Mittelwerttests

Bei den klassischen Mittelwerttests handelt es sich um statistische Hypothesentests mit einer einfachen Alternativen. Die Hypothese begründet sich auf der Annahme, daß sich Geschwister an jedem beliebigen Genort im Durchschnitt etwa die Hälfte ihrer Allele identisch nach Herkunft teilen. Um systematische Abweichungen vom erwarteten mittleren Anteil Allele IBD festzustellen, schlugen Green und Woodrow (1977) einen *Mittelwerttest* (engl.: mean test) vor. Es ist zu beachten, daß dieser klassische Ansatz bei der Bestimmung des Anteils Allele IBD eines Geschwisterpaares davon ausgeht, daß sich IBD-Werte eindeutig bestimmen lassen. Konsequenterweise wurde dann auch die theoretische Varianz des Anteils der Allele IBD zur Berechnung verwendet. Ein etwas allgemeineres und realistischeres Modell welches sowohl die IBD-Verteilung berücksichtigt, als auch eine empirische Bestimmung der Varianz aus den Daten zuläßt, wurde von deVries et al. (1976) vorgeschlagen. Blackwelder und Elston (1985) verglichen die Power dieser beiden Ansätze und fanden, daß der allgemeinere Ansatz mit den meisten genetischen Modellen eine höhere statistische Güte als der klassische besitzt. In einigen Fällen eines rezessiven genetischen Modells ist der *Anteilstest* (engl.: proportion test) (Day und Simons, 1976) geeigneter eine Abweichung von der Erwartung festzustellen. Der Anteilstest betrachtet nicht den durchschnittlichen Anteil der Allele IBD, sondern die durchschnittliche Wahrscheinlichkeit eines Paares, sich genau zwei Allele IBD zu teilen. Dabei beträgt die a-priori Erwartung für den durchschnittlichen IBD-Wert beim Anteilstest ein Viertel.

Obwohl der klassische Mittelwerttest bereits gewisse Optimalitätseigenschaften besitzt (Knapp et al., 1994), kann er noch weiter verbessert werden. Als Beispiel sei eine Stichprobe von 20 Geschwisterpaaren mit jeweils einem IBD-Wert von 2 angeführt. Der Mittelwerttest von Green und Woodrow (1977) ergibt hier einen p-Wert von $1.27 \cdot 10^{-10}$ (einen LOD-Score von 8.68). Fügt man dieser Stichprobe 20 nicht-informative Geschwisterpaare hinzu, so sollten diese das Ergebnis aufgrund ihrer nicht-Informativität nicht verändern. Der mit der vergrößerten Stichprobe berechnete Mittelwerttest ergibt aber einen p-Wert von $3.87 \cdot 10^{-6}$ (oder einen LOD-Score von 4.34). Dieses Beispiel zeigt, daß uninformative Geschwisterpaare keinen Gewinn für die Studie darstellen, im Gegenteil, sie verringern die Chance einen signifikanten Zusammenhang zwischen Genort und Krankheit bei den anderen Probanden zu entdecken.

Die Begründung für diesen, auf den ersten Blick vielleicht überraschenden Sachverhalt ist sehr einfach. Obwohl die IBD-Schätzungen und damit auch der Anteil der Allele IBD eine deutliche Variabilität bezüglich ihrer Informativität aufweisen (vgl. Abschnitt 2.2.4), fließen alle Beobachtungen, auch die nicht-informativen, mit dem gleichen Gewicht in die Analyse ein. Wie allerdings von Horvitz und Thompson (1952) im Zusammenhang mit allgemeinen Verfahren der Stichprobentheorie ausgeführt, erscheint es sinnvoller die Beobachtungen anhand ihrer Informativität zu gewichten.

Dabei ist zu beachten, daß Geschwisterpaare nicht anhand ihrer Phänotypen gewichtet werden können, nach Design sind beide Probanden erkrankt. Der Ansatz Geschwisterpaaren über ihre Genotypinformationen Gewichte zuzuweisen war bisher erfolglos. In dieser Arbeit wird ab Abschnitt 4.2 ein, auf

IBD-Informationen basierender, gewichteter Mittelwerttest beschrieben. Der folgende Abschnitt faßt bisherige Gewichtungsansätze bei Mittelwerttests zusammen.

4.1.2 Mittelwerttests mit Gewichtung

Modellfreie Verfahren, wie die hier betrachteten, legen keine Annahmen über die Vererbungsmuster zugrunde. Trotzdem beeinflußt das Vererbungsmodell die Wahrscheinlichkeit Kopplung zu entdecken (Blackwelder und Elston, 1985). Diesem Umstand trägt der *minmax-Test* von Whittemore und Tu (1998) Rechnung: da der Mittelwerttest bei dominanten Vererbungsmodellen gute Ergebnisse zeigt und der Anteilstest eher bei rezessiver Vererbung, so argumentieren Whittemore und Tu, muß es eine Parametrisierung dieser Tests geben, so daß die statistische Güte für alle Vererbungsmodelle zumindest akzeptabel bleibt. Whittemore und Tu verwenden die Mittelwertsstatistik als Grundlage und führen einen zusätzlichen Gewichtungsparameter ein, derart, daß, je nach Wahl dieses Parameters, genau der Mittelwerttest, genau der Anteilstest oder aber ein intermediärer Test generiert wird. Der resultierende Test, dessen Parameter eine optimale Gewichtung zwischen Mittelwert- und Anteilstest ermöglicht, wird minmax-Test genannt.

Eine andere Problemstellung wurde von Sham et al. (1997) untersucht. Die Autoren betrachteten, wie unterschiedlich große Familien mit einer variierenden Anzahl von Geschwistern miteinander zu kombinieren wären. Wie Suarez und Hodge (1979) und Hodge (1984) ausführten, ist die Informativität mehrerer Geschwisterpaare in einer Familie nicht gleichzusetzen mit

der Informativität mehrerer Familien mit je einem Geschwisterpaar. Unter Verwendung des Shannon'schen Entropiemaßes bestimmten Suarez und Hodge (1979) die Information, die in einer Familie mit s Nachkommen gegeben ist. Sie zeigten, daß alle möglichen Geschwisterpaarungen gemeinsam soviel Information beitragen wie $(2s - 3 + 0.5^{s-1})/1.5$ unabhängige Paare. Eine aus 4 Nachkommen bestehende Geschwisterschaft ist trotz der insgesamt 6 unterschiedlichen Paare also nur so informativ wie etwa 3.4 unabhängige Geschwisterschaften. Trotzdem kamen Sham et al. (1997) zu dem Ergebnis, daß man im Allgemeinen, solange weniger als fünf erkrankte Geschwister in einer Familie betrachtet werden, alle Familien gleich behandelt werden können. Eine Gewichtung der einzelnen Familien zueinander ist somit nicht erforderlich (vgl. Blackwelder und Elston, 1985).

4.2 Methoden

Die bisher vorgestellten Teststatistiken nach Green und Woodrow (1977) und deVries et al. (1976) lassen sich in einer allgemeinen Form darstellen:

$$T = \frac{\bar{\tau} - \mathbb{E}_{H_0}(\bar{\tau})}{\sqrt{\mathbb{V}ar(\bar{\tau})}} \quad .$$

Je nach Definition der Mittelwertsfunktion $\bar{\tau}$ bzw. der Definition der Varianz der Mittelwertsfunktion, $\mathbb{V}ar(\bar{\tau})$, ergeben sich verschiedene Score bzw. Wald Tests. Aus einer speziellen Definition dieser Funktionen wird in den nächsten Abschnitten der gewichtete Mittelwerttest hergeleitet werden. Der Erwartungswert des mittleren Anteils Allele IBD τ ist unter der Nullhypo-

these für alle Tests einheitlich. Er entspricht der impliziten Annahme, daß sich Geschwisterpaare an Genorten ohne Kopplung im Durchschnitt ein Allel identisch nach Herkunft teilen, d. h. $\mathbb{E}_{H_0}(\hat{\tau}) = \frac{1}{2}$.

4.2.1 Klassische Mittelwerttests

Werden n unabhängige Geschwisterpaare betrachtet, so ist $n = n_0 + n_1 + n_2$, wobei n_i $(i = 0, 1, 2)$ die Anzahl derjenigen Paare sei, die sich an dem betrachteten Genort genau i Allele IBD teilen. Hierbei wird die eindeutige Bestimmbarkeit der IBD-Werte vorausgesetzt.

Für den klassischen Mittelwerttest sind die Mittelwertsfunktion sowie ihre Varianz nach Green und Woodrow (1977) geschätzt durch:

$$\bar{\hat{\tau}} = \frac{n_2}{n} + \frac{n_1}{2n} \qquad\qquad \mathbb{V}ar\left(\bar{\hat{\tau}}\right) = \frac{1}{8n} \quad .$$

Der Score-Test

$$T_m = \frac{\left(\frac{n_2}{n} + \frac{n_1}{2n}\right) - \frac{1}{2}}{\sqrt{\frac{1}{8n}}} \tag{4.1}$$

folgt unter der Nullhypothese einer Standardnormalverteilung. Unter der Alternativen verschiebt sich der Mittelwert der Verteilung auf einen von 0 verschiedenen Wert.

Läßt man die Annahme der eindeutig definierten IBD-Werte zugunsten von Wahrscheinlichkeitsverteilungen fallen, so werden der Mittelwertsschätzer durch das arithemtische Mittel der Anteile Allele IBD der Geschwisterpaare und die analytische Varianz der IBD-Werte durch die empirische

Varianz der Anteile Allele IBD ersetzt (deVries et al., 1976):

$$\bar{\hat{\tau}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i \qquad \widehat{\mathbb{V}ar}\left(\bar{\hat{\tau}}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\tau}_i - \bar{\hat{\tau}}\right)^2$$

Der Wald-Test

$$T_{m,ev} = \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i - \frac{1}{2}}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\tau}_i - \bar{\hat{\tau}}\right)^2}} \tag{4.2}$$

folgt unter der Nullhypothese einer t-Verteilung mit $n-1$ Freiheitsgraden.

4.2.2 Gewichteter Mittelwertwerttest

Der im Folgenden beschriebene Ansatz eines gewichteten Mittelwerttest wurde erstmals von Franke und Ziegler (2005) vorgestellt.

Teststatistik

Für den gewichteten Mittelwerttest wird das arithmetische durch ein gewichtetes Mittel ersetzt, die Varianzfunktion durch eine gewichtete emprische Varianz:

$$\bar{\hat{\tau}} = \sum_{i=1}^{n} w_i\hat{\tau}_i \qquad \widehat{\mathbb{V}ar}\left(\bar{\hat{\tau}}\right) = \sum_{i=1}^{n} w_i^2\left(\hat{\tau}_i - \bar{\hat{\tau}}\right)^2 \qquad \text{wobei} \qquad \sum_{i=1}^{n} w_i = 1 \quad .$$

Es sei n' die Anzahl derjenigen Beobachtungen, deren Gewicht sich von null unterscheidet, also $1 < n' \leq n$. Dann ist der gewichtete Mittelwerttest T_w

definiert als

$$T_w = \frac{\sum_{i=1}^{n} w_i \hat{\tau}_i - \frac{1}{2}}{\sqrt{\frac{n'}{n'-1} \sum_{i=1}^{n} w_i^2 \left(\hat{\tau}_i - \bar{\hat{\tau}}\right)^2}} \quad , \tag{4.3}$$

wobei der Vorfaktor $\frac{n'}{n'-1}$ zur Varianz sicherstellt, daß T_w in $T_{m,ev}$ übergeht, wenn $w_i = \frac{1}{n}$ gewählt wird.

Asymptotische Verteilung

Die asymptotische Verteilung ist aufgrund der frei wählbaren Gewichte w_i nicht offensichtlich. Verallgemeinerte Schätzgleichungen (engl.: generalised estimating equations, GEE) können hier Hilfestellung leisten (Ziegler et al., 1998). Wird in einem GEE-Modell die Einheitsmatrix als Arbeitskorrelationsmatrix gewählt, so kann der gewichtete Mittelwerttest aus diesem Modell heraus abgeleitet werden. Hierzu wird der Anteil Allele IBD τ_i der n Geschwisterpaare additiv in einen wahren, gemeinsamen Wert τ und einen geschwisterspezifischen Fehlerterm ε_i zerlegt. Es gelte desweiteren $\mathbb{E}(\tau_i) = \tau$ und $\mathbb{V}ar(\tau_i) = w_i$. Schätzt man τ unter Verwendung eines gewichteten Kleinste Quadrate Schätzers mit Gewichtsmatrix $\mathrm{diag}(w_i)$, so kann man unter zusätzlicher Anwendung des robusten Varianzschätzers (Ziegler et al., 1998) den in Gleichung (4.3) beschriebenen Test ableiten. Da die asymptotischen Eigenschaften des GEE-Schätzers bekannt sind, lassen sie sich auf den vorliegenden Fall anwenden: unter der Nullhypothese, daß keine Kopplung vorliegt, folgt T_w asymptotisch einer Standardnormalverteilung.

Für kleine Fallzahlen kann anstelle einer Standardnormalverteilung eine t-Verteilung mit $n' - 1$ Freiheitsgraden angenommen werden, um zu liberale Testergebnisse zu vermeiden.

Gewichtungsschemata für den gewichteten Mittelwerttest

Analog zur Definition der Gewichtungsschemata für das Verfahren von Haseman und Elston (vgl. Abschnitt 3.2.3) wird auch hier von einem Informationsmaß auf ein Gewichtungsschema geschlossen. Prinzipiell sind hierzu alle in Abschnitt 2.2.5 aufgeführten Informationsmaße geeignet, exemplarisch wird hier der euklidische Abstand herausgegriffen.

Der Abstand einer IBD-Schätzung v von der uninformativen a-priori Schätzung für Geschwisterpaare u, ist in der Ebene des de Finetti-Diagrams durch Gleichung (2.2) aus Abschnitt 2.2.5 gegeben:

$$d_{euklid}\left(v_0, v_1, v_2\right) = \left(\frac{\left(v_0 - v_2\right)^2}{3} + \frac{\left(1 - 2v_1\right)^2}{4} \right)^{-\frac{1}{2}} \quad .$$

Die eigentlichen Gewichte w_i werden durch Normierung der Abstände $d_{euklid}(v_i)$ so bestimmt, daß $\sum_{i=1}^{n} w_i = 1$ gilt:

$$w_i(v_i) = \frac{d_{euklid}(v_i)}{\sum_{j=0}^{n} d_{euklid}(v_j)} \quad . \tag{4.4}$$

4.2.3 Ermittlung empirischer p-Werte durch Monte-Carlo Simulation

Da zu Beginn dieser Arbeit nicht absehbar war, ob sich überhaupt eine asymptotische Verteilung des gewichteten Mittelwerttests herleiten ließe, wurden auch Möglichkeiten untersucht, p-Werte über Permutations- oder Simulationsansätze zu gewinnen. Es können einige bisher bekannte Permutationsverfahren unterschieden werden:

Sowohl in Kopplungsanalysen quantitativer Phänotypen mit Geschwisterpaaren als auch bei Fall-Kontroll Studien können wahlweise die Phänotypen oder die Genotypen permutiert werden, um eventuelle Abhänigkeiten aufzulösen (Wan et al., 1997; Zhao et al., 2000). Dies ist in Studien mit erkrankten Geschwisterpaaren nicht möglich, da alle Probanden den gleichen Phänotypstatus tragen.

Bei vorgegebener Heterozygotie könnten neue Markergenotypen unter der Nullhypothese generiert werden (Zinn-Justin et al., 2001). Unglücklicherweise würde dies für jede Wiederholung zu einem neuen Set von Gewichten führen, der Informationsinhalt der simulierten Stichprobe wäre für jede Iteration ein anderer.

Das von Zhao et al. (1999) vorgeschlagene Randomisierungsverfahren für die Vererbungsvektoren des Lander-Green Algorithmus (Idury und Elston, 1997; Kruglyak und Lander, 1998) läßt sich auf die vorliegende Situation ebenfalls nicht anwenden. Auch hier würde sich die enthaltene Informativität und damit die zugeordneten Gewichte verändern.

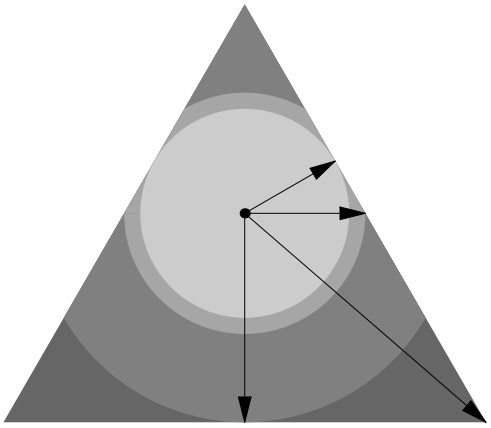


Abbildung 4.1: Diese Abbildung verdeutlicht die in Tab. 4.1 angegebenen Fallunterscheidungen. Der hellgraue, innere Bereich entspricht einer Kreisscheibe um den Punkt der Nichtinformativität – also $\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$. Die dunkler werdenden Bereiche ergeben sich durch die Begrenzungen der Dreiecksseiten. Der äußerste Bereich deckt die Restfläche bis zu den Ecken des Dreiecks ab und ist dunkelgrau gekennzeichnet.

Da im vorliegenden Fall keines der etablierten Verfahren angewendet werden konnte, wird im folgenden eine neue Methode vorgestellt.

Permutation von IBD-Werten bei gleichbleibender Informativität

Zwei IBD-Werte werden als gleich informativ angesehen, wenn sie unter einer Definition eines Abstandsbegriffes die gleiche Entfernung zur uninformativen IBD-Schätzung aufweisen, d. h. im Falle des zuvor beschriebenen euklidischen Abstandes, wenn sie auf kreisförmigen Isolinien um den uninformativen Wert $u = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ liegen. Es ist zu beachten, daß für eine andere Abstandsdefinition möglicherweise nicht-kreisförmige Isolinien entstehen können: durch Anwendung der Maximumbetragsnorm ergeben sich hexagon- statt kreisförmige Isolinien. Im weiteren wird von den kreisförmigen Isolinien des euklidischen Abstandsbegriffes ausgegangen.

Mit wachsender Informativität geht der anfängliche Vollkreis in eine Vereinigung von Kreisbögen über, die schließlich auf zwei Punkte maximaler Information degenerieren. Aufgrund dieser Struktur ist es sinnvoll gültige

Fall	Abstand d	Gültiger Bereich von t
1	$0 \leq d \leq \frac{1}{4}$	$0 \leq t < 2\pi$
2	$\frac{1}{4} < d < \frac{1}{2\sqrt{3}}$	$0 < t \leq 2\arctan\left(3^{-\frac{1}{2}} - a - A\right)$
		$2\arctan\left(3^{-\frac{1}{2}} - a + A\right) \leq t \leq 2\arctan\left(-3^{-\frac{1}{2}} + b - B\right)$
		$-2\arctan\left(3^{-\frac{1}{2}} - b - B\right) \leq t < \pi$
		$\pi \leq t \leq 2\pi$
3	$d = \frac{1}{2\sqrt{3}}$	$\frac{\pi}{3} \leq t \leq \frac{2\pi}{3}$
		$\pi \leq t \leq 2\pi$
4	$\frac{1}{2\sqrt{3}} < d \leq \frac{1}{2}$	$2\arctan\left(3^{-\frac{1}{2}} - a + A\right) \leq t \leq 2\arctan\left(-3^{-\frac{1}{2}} + b + B\right)$
		$2\pi - 2\arctan\left(3^{-\frac{1}{2}} - b + B\right) \leq t \leq 2\pi - 2\arctan\left(-3^{-\frac{1}{2}} + a + A\right)$
5	$\frac{1}{2} < d \leq \sqrt{\frac{7}{12}}$	$2\pi - 2\arctan\left(3^{-\frac{1}{2}} - b + B\right) \leq t \leq 2\pi - 2\arctan\left(2d + \sqrt{4d^2 - 1}\right)$
		$2\pi - 2\arctan\left(2d - \sqrt{4d^2 - 1}\right) \leq t \leq 2\pi - 2\arctan\left(-3^{-\frac{1}{2}} + a + A\right)$

Wobei

$$a = \left(\sqrt{3} + 6d\right)^{-1} \qquad A = \sqrt{a^2\, 3(16d^2 - 1)}$$
$$b = \left(\sqrt{3} - 6d\right)^{-1} \qquad B = \sqrt{b^2\, 3(16d^2 - 1)}$$

Tabelle 4.1: Resampling-Vorschriften zur Bestimmung empirischer p-Werte. Die Berechnung gültiger IBD-Werte basiert auf Polarkoordinaten (d, t) mit Ursprung $(f_0, f_1, f_2) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. Zu einem euklidischen Abstand d sind hier die zugehörigen gültigen Winkel t in rad angegeben, so daß jedes Paar (d, t) sich auf einen gültigen IBD-Wert abbilden läßt.

Bereiche für IBD-Werte in Vereinigungen von Paaren von Polarkoordinaten der Form $\{(d, t_{Anfang}), (d, t_{Ende})\}$ anzugeben. Die entstehende Vereinigungsmenge quantifiziert den Abstand d sowie die Radianten des Anfangs bzw. des Endes eines gültigen Bereiches. Abbildung 4.1 zeigt farblich die Bereiche der notwendigen Fallunterscheidungen bei der Berechnung des zulässigen Winkelbereiches für t. Sämtliche Fallunterscheidungen sind in Tab. 4.1 zusammengefasst.

Nimmt man die in Tab. 4.1 dargestellten Fälle als Grundlage, läßt sich ein einfacher Algorithmus zur Permutation von IBD-Werten bei gleichbleibender Informativität im Sinne des euklidischen Abstandes angeben:

1. Berechne die gewichtete Teststatistik T_w aus den Originaldaten
2. Wiederhole für jedes Geschwisterpaar $1 \leq i \leq n$:
   (a) Ziehe aus einer Gleichverteilung über dem in Tab. 4.1 für d_i angegebenen Bereich einen zufälligen Wert für t_i
   (b) Berechne die neue IBD-Verteilung nach der Vorschrift

$$f_{0i} = \frac{1}{4} + \frac{d_i}{2}\left(\sqrt{3}\cos(t_i) - \sin(t_i)\right)$$
$$f_{1i} = \frac{1}{2} + d_i \sin(t_i)$$
$$f_{2i} = 1 - f_{0i} - f_{1i}$$

3. Berechne die Teststatistik T_{sim} aus den generierten Daten
4. Wiederhole Schritte 2. und 3. M-mal
5. Berechne den empirischen p-Wert über $\#\{T_{sim} > T_w\}/M$

Die Wahl von M beeinflußt die Präzision des ermittelten empirischen p-Wertes. Ein typischer Wert für Simulationen unter der Nullhypothese wäre beispielsweise $M = 100.000$, für Simulationen unter der Alternativen $M = 1.000$ (vgl. Abschnitt 2.3.3).

Zur Verdeutlichung des Punkts 2.a) des Algorithmus: die geschätzte IBD-Verteilung sei $v = (0.5, 0.1, 0.4)$, dann beträgt der euklidische Abstand zur a-priori Schätzung $d = \frac{7\sqrt{3}}{30}$, dies entspricht Fall 4 aus Tab. 4.1, da $d \in \left(\frac{1}{2\sqrt{3}}; \frac{1}{2}\right]$. Mit

$$a = \frac{5\sqrt{3}}{36} \approx 0.2406, \qquad A = \frac{11\sqrt{3}}{36} \approx 0.5292,$$
$$b = \frac{5\sqrt{3}}{6} \approx -1.4434, \qquad B = \frac{11\sqrt{3}}{6} \approx 3.1754,$$

muß t aus einer stetigen Gleichverteilung über

$$\left[2\arctan\left(\frac{\sqrt{3}}{2}\right); -2\arctan\left(\frac{2\sqrt{3}}{3}\right)\right] \cup \left[2\pi - 2\arctan\left(3\sqrt{3}\right); 2\pi - 2\arctan\left(\frac{\sqrt{3}}{9}\right)\right]$$
$$\approx [1.4275; 1.7141] \cup [3.5218; 5.9029]$$

gezogen werden.

4.3 Simulationsstudien

In diesem Abschnitt soll mittels Simulationsstudien sowohl die Zulässigkeit als auch die Validität des gewichteten Mittelwerttests im Sinne des Fehlers 1. Art bzw. im Sinne der statistischen Güte gezeigt werden.

Zur Simulation der Geschwisterpaare wurden anstelle des vollständigen Segregationspfades, also die zufällige Ziehung der Allele der Eltern, Segregation der Allele an die Nachkommen, Bestimmung des Krankheitszustandes und der IBD-Verteilung, die zu erwartenden Häufigkeiten der sieben möglichen IBD-Verteilungen unter Annahme bestimmter genetischer Modelle herangezogen (Tab. 4.2). Durch diese Abkürzung ließen sich sowohl Zeit- als auch Rechenaufwand für die Simulation deutlich verringern.

Drei verschiedene genetische Modelle wurden betrachtet: der ungekoppelte Fall, ein autosomal dominantes sowie ein „freies“ Vererbungsmodell. Im ungekoppelten Fall wurde ein Marker mit r gleich häufigen Allelen angenommen (vgl. Risch, 1990). Sowohl beim dominanten als auch beim freien Modell wurde eine vollständige Penetranz ohne Phänokopien zugrunde gelegt. Dabei sei die Krankheitsallelfrequenz p, die Rekombinationswahrscheinlichkeit $\theta = 0$ und der Markerlocus habe wie beim ungekoppelten Modell r gleichwahrscheinliche Allele. Desweiteren wurde angenommen, daß beide Marker sich im Hardy-Weinberg-Gleichgewicht (Ziegler und König, 2006, Kap. 2.4) befinden und kein Kopplungsungleichgewicht (Ziegler und König, 2006, Kap. 9.2) vorliegt. Die Wahrscheinlichkeiten für das dominante sowie ein rezessives Modell sind in Tab. 4.2 zusammengestellt. Die Wahrscheinlichkeiten für das freie Modell wurden in Franke und Ziegler (2005, Tab. 2) angegeben

	Erwartete Auftretenswahrscheinlichkeiten		
IBD Wert	Dominantes Modell	Rezessives Modell	Freies Modell
$(1,0,0)$	$\frac{r^3-2r^2+1}{4r^3}\left(p^4-4p^3+4p^2\right)$	$\frac{r^3-2r^2+1}{4r^3}\,p^4$	$\frac{r-1}{8r^3}\left((3r^2-3r-4)p^4-(4r^2-4r-6)p^3+(3r^2-3r-4)p^2\right)$
$(0,1,0)$	$\frac{(r-1)^2}{2r^2}\left(-p^3+p^2+p\right)$	$\frac{(r-1)^2}{2r^2}\,p^3$	$\frac{(r-1)^2}{8r^2}\left(5p^4-5p^3+3p^2+p\right)$
$(0,0,1)$	$\frac{r^3-2r^2+1}{4r^3}\left(-p^2+2p\right)$	$\frac{r^3-2r^2+1}{4r^3}\,p^2$	$\frac{1}{16r^3}\big((5r^3-10r^2-r+6)p^4-(4r^3-8r^2-2r+6)p^3$ $+(r^3-2r^2-r+2)p^2+(2r^3-4r^2+2)p\big)$
$(\frac{1}{2},\frac{1}{2},0)$	$\frac{r-1}{r^2}\,\frac{p^4-5p^3+5p^2+p}{2}$	$\frac{r-1}{r^2}\,\frac{p^4+p^3}{2}$	$\frac{r-1}{4r^2}(5p^4-5p^3+3p^2+p)$
$(\frac{1}{2},0,\frac{1}{2})$	$\frac{r-1}{2r^3}\,\frac{p^4-4p^3+3p^2+2p}{2}$	$\frac{r-1}{2r^3}\,\frac{p^4+p^2}{2}$	$\frac{r-1}{8r^3}(7p^4-9p^3+5p^2+p)$
$(0,\frac{1}{2},\frac{1}{2})$	$\frac{r-1}{r^2}\,\frac{-p^3+3p}{2}$	$\frac{r-1}{r^2}\,\frac{p^3+p^2}{2}$	$\frac{r-1}{8r^2}(11p^4-12p^3+7p^2+2p)$
$(\frac{1}{4},\frac{1}{2},\frac{1}{4})$	$\frac{1}{r^2}\,\frac{p^4-6p^3+5p^2+4p}{4}$	$\frac{1}{r^2}\,\frac{p^4+2p^3+p^2}{4}$	$\frac{1}{16r^2}(21p^4-22p^3+13p^2+4p)$

Tabelle 4.2: Auftretenswahrscheinlichkeiten von IBD-Verteilungen bei zwei erkrankten Nachkommen in einer Kernfamilie unter verschiedenen genetischen Modellen mit r gleichhäufigen Markerallelen (Suarez et al., 1978; Risch, 1990). Der ungekoppelte Fall wurde bereits in Tabelle 3.2 dargestellt. Im dominanten, rezessiven bzw. . freien Modell gilt eine Rekombinationswahrscheinlichkeit $\theta = 0$. Desweiteren wird für alle Modelle volle Penetranz ohne Phänokopien vorausgesetzt, dabei p sei die Wahrscheinlichkeit des Krankheitsalleles am biallelischen Krankheitsgenort. Es wird angenommen, daß sich die Marker im Hardy-Weinberg-Gleichgewicht (vgl. Ziegler und König, 2006, Kap. 2.4) befinden und es kein Kopplungsungleichgewicht gibt (vgl. Ziegler und König, 2006, Kap. 9.2).

(dort allerdings fälschlicherweise unter der Annahme ein rezessives Modell zu simulieren, dies wurde im Erratum zu Franke und Ziegler (2005) korrigiert).

Die für die Simulation veränderlichen Parameter waren die Wahrscheinlichkeit des Krankheitsalleles ($p = 0.00001$), die Anzahl der Familien ($n = 50$ und $n = 200$), sowie die Anzahl der gleichhäufig vorkommenden Allele am Markerlocus: $r = 2$ entspricht hierbei in etwa einer Heterozygotie von 50%, was mit einer gängigen Zweipunktanalyse vergleichbar wäre, $r = 4$ entspricht einer Heterozygotie von 75% und ist damit mit einem üblichen Mikrosatelliten-Scan gleichzustellen. Ein mit $r = 100$ gleichhäufigen Allelen belegter Marker entspricht einer fast voll informativen Position im Chromosom und kann beispielsweise in genomweiten SNP-Scans (engl.: Single Nucleotide Polymorphism, SNP) erreicht werden. Mehr Informationen über STR- und SNP-Marker können beispielsweise Strachan und Read (2004) entnommen werden.

Unter der Nullhypothese, daß keine Kopplung besteht, wurden für jede Kombination aus Anzahl Familien n und Anzahl Allelen r je 100.000 Datensätze simuliert; unter der Alternativen des dominanten Modells und vollständiger Kopplung nochmals weitere 1.000. Aysmptotische p-Werte wurden sowohl für den klassischen als auch für den gewichteten Mittelwerttest über die t-Verteilung mit $n-1$ bzw. $n'-1$ Freiheitsgraden bestimmt. Empirische p-Werte wurden über den zuvor beschriebenen, neuartigen, Simulationsansatz ermittelt. Dazu wurde die Anzahl der Replikationen auf $M = 10.000$ gesetzt. Zur Berechnung der klassischen Teststatistik wurden die euklidischen Abstände ignoriert und die Gewichte auf $\frac{1}{n}$ festgelegt. Die zu einem Signifikanzniveau α gehörende statistische Güte wurde über das obere α-Fraktil

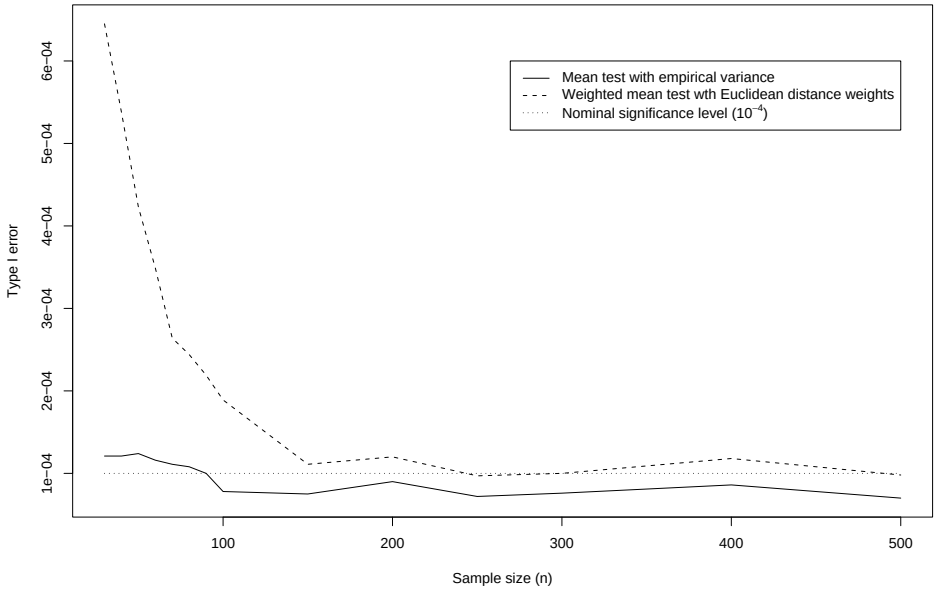


Abbildung 4.2: Vergleich Fehler 1. Art des klassischen und des gewichteten Mittelwerttests bei wachsendem Stichprobenumfang n. Für jede Fallzahl n wurden 1.000.000 Datensätze unter der Nullhypothese generiert, die p-Werte wurden einer $t-$Verteilung mit $n-1$ Freiheitsgraden entnommen, das Signifikanzniveau beträgt $\alpha = 10^{-4}$.

der asymptotischen bzw. der empirischen Verteilung der p-Werte unter der Nullhypothese bestimmt. Alle Ergebnisse wurden mit Hilfe eines in C++ geschriebenen Simulationsprogrammes ermittelt.

Die Ergebnisse der Simulationen sind nach Fehlerhäufigkeiten 1. Art und statistischer Güte getrennt in den Tabellen 4.3 und 4.4 aufgeführt. Wie in Abb. 4.2 zu sehen ist, halten die auf der t-Verteilung beruhenden p-Werte das Signifikazniveau von $\alpha = 0.0001$ ab etwa $n = 100$ Beobachtungen sehr gut ein. Für den neuen, gewichteten Ansatz etwas später (gestrichelte Linie in Abb. 4.2) als für den Mittelwertstest (durchgezogene Linie). Die Begründung hierfür ist wahrscheinlich in den zusätzlichen zufälligen Elementen, die von

	Fehler 1. Art											
	$n = 50, r = 2$		$n = 50, r = 4$		$n = 50, r = 100$		$n = 200, r = 2$		$n = 200, r = 4$		$n = 200, r = 100$	
Nominaler α	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w
Asymptotisch												
0.001	0.0011	0.0015	0.0010	0.0020	0.0010	0.0018	0.0009	0.0013	0.0008	0.0009	0.0010	0.0010
0.01	0.0099	0.0138	0.0103	0.0136	0.0097	0.0120	0.0095	0.0102	0.0096	0.0101	0.0101	0.0102
0.02	0.0205	0.0260	0.0197	0.0240	0.0212	0.0220	0.0197	0.0205	0.0194	0.0195	0.0204	0.0203
0.03	0.0306	0.0376	0.0300	0.0341	0.0313	0.0320	0.0297	0.0306	0.0292	0.0298	0.0309	0.0302
0.04	0.0409	0.0487	0.0403	0.0443	0.0401	0.0425	0.0399	0.0412	0.0391	0.0397	0.0408	0.0401
0.05	0.0503	0.0599	0.0491	0.0546	0.0498	0.0532	0.0502	0.0514	0.0491	0.0500	0.0511	0.0506
Empirisch												
0.001	0.0012	0.0005	0.0008	0.0008	0.0010	0.0013	0.0012	0.0010	0.0010	0.0007	0.0009	0.0010
0.01	0.0107	0.0082	0.0103	0.0100	0.0096	0.0104	0.0106	0.0089	0.0101	0.0096	0.0103	0.0101
0.02	0.0203	0.0176	0.0198	0.0199	0.0200	0.0204	0.0212	0.0185	0.0203	0.0189	0.0206	0.0203
0.03	0.0304	0.0281	0.0295	0.0298	0.0297	0.0303	0.0313	0.0285	0.0302	0.0288	0.0309	0.0300
0.04	0.0404	0.0381	0.0399	0.0397	0.0401	0.0403	0.0415	0.0388	0.0401	0.0391	0.0413	0.0396
0.05	0.0503	0.0486	0.0492	0.0500	0.0500	0.0508	0.0517	0.0493	0.0506	0.0491	0.0515	0.0501

Tabelle 4.3: Vergleich der Fehlerhäufigkeiten zwischen $T_{m,ev}$ und T_w (euklidische Abstandsgewichte), die Fehlerraten sind angegeben für n Paare bei r gleichhäufigen Markerallelen. Asymptotische p-Werte wurden über die t-Verteilung mit $n - 1$ Freiheitsgraden bestimmt, empirische p-Werte über den neuen Permutationsansatz.

	Statistische Güte, Dominantes Modell				Statistische Güte, Freies Modell			
	$n = 200, r = 2$		$n = 200, r = 4$		$n = 200, r = 2$		$n = 200, r = 4$	
Nominaler α	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w	$T_{m,ev}$	T_w
Asymptotisch								
0.001	0.998	1.000	1.000	1.000	0.156	0.594	1.000	1.000
0.01	0.999	1.000	1.000	1.000	0.398	0.864	1.000	1.000
0.02	1.000	1.000	1.000	1.000	0.498	0.920	1.000	1.000
0.03	1.000	1.000	1.000	1.000	0.561	0.946	1.000	1.000
0.04	1.000	1.000	1.000	1.000	0.607	0.956	1.000	1.000
0.05	1.000	1.000	1.000	1.000	0.649	0.964	1.000	1.000
Empirisch								
0.001	0.998	1.000	1.000	1.000	0.167	0.559	1.000	1.000
0.01	0.999	1.000	1.000	1.000	0.405	0.862	1.000	1.000
0.02	1.000	1.000	1.000	1.000	0.498	0.918	1.000	1.000
0.03	1.000	1.000	1.000	1.000	0.565	0.946	1.000	1.000
0.04	1.000	1.000	1.000	1.000	0.607	0.956	1.000	1.000
0.05	1.000	1.000	1.000	1.000	0.648	0.964	1.000	1.000

Tabelle 4.4: Vergleich der statistischen Güte von $T_{m,ev}$ und T_w (euklidische Abstandsgewichte), die Güte ist angegeben für 200 Paare bei $r = 2, 4$ gleichhäufigen Markerallelen. Zur Simulation der Paare wurde zum einen ein autosomal dominantes genetisches Zweipunktmodell mit voller Penetranz ohne Phänokopien und $\theta = 0$ zugrunde gelegt. Da der gewichtete Test für kleine Fallzahlen ($n = 50$) zur Liberalität neigt, werden die zugehörigen Werte hier nicht dargestellt. Zum anderen stellt das freie Modell Ergebnisse dar, wie sie bereits in Franke und Ziegler (2005) veröffentlicht wurden. Sie basieren auf den dort, von Tabelle 4.2 abweichenden, Wahrscheinlichkeiten (Franke und Ziegler, 2005, Tab. 2).

den einzelnen w_i eingeführt werden, zu suchen. Wie Abb. 4.2 zeigt, sind die Unterschiede vernachlässigbar, sobald die Anzahl der Beobachtungen $n = 150$ übersteigt. Die empirischen p-Werte halten das Niveau für jedes Niveau α (vgl. Tab. 4.3).

Wie aus Tab. 4.3 zu ersehen ist, fällt der gewichtete Test bei kleinen Fallzahlen ($n = 50$) zu liberal aus, daher wurde auf eine Angabe der zugehörigen Werte für die statistische Güte in Tab. 4.4 verzichtet. Tabelle 4.4 zeigt, steigende Güte bei steigender Anzahl der Markerallele bzw. bei steigender Fallzahl. Es besteht nur ein geringer Unterschied zwischen asymptotischer oder empirischer Bestimmung der statistischen Güte. Im Freien Modell ist ein bemerkenswerter Unterschied der Güte zwischen den beiden Tests, dem klassischen Ansatz mit empirischer Varianz und dem neuen gewichteten Ansatz, zu erkennen. Beispielsweise läßt sich bei einem Signifikanzniveau von $\alpha = 0.001$, einer niedrigen Heterozygotie von 50% ($r = 2$) und $n = 200$ Fällen, ein Anstieg der Wahrscheinlichkeit Kopplung zu entdecken, von 15% auf 60% verzeichnen.

Es bleibt festzustellen, daß die hier vorgestellten Simulationsstudien nur einen kleinen Teil aller möglichen Fälle beleuchten. Nach momentanem Kenntnisstand sind Situationen, in denen der gewichtete Test auch schlechter als der klassische Ansatz abschneiden könnte, nicht vollständig auszuschließen.

4.4 Anwendungen

Anhand von zwei Beispielen soll die Anwendung des gewichteten Mittelwerttests verdeutlicht werden.

Anzahl der Paare	f_2	f_1	f_0
14	1	0	0
16	0	1	0
2	0	0	1
22	$\frac{1}{2}$	$\frac{1}{2}$	0
13	0	$\frac{1}{2}$	$\frac{1}{2}$
7	$\frac{1}{2}$	0	$\frac{1}{2}$

Tabelle 4.5: Datensatz zum Vergleich der Teststatistiken in Tab. 4.6 (vgl. Risch, 1990)

Als erstes wird hierzu das von Risch (1990, Tab. 1) eingeführte Datenmaterial übernommen, welches von Risch (1990) verwendet wurde um die Maximum-LOD-Score (MLS) Statistik zu verdeutlichen. Außerdem werden die Daten der Studie zur Diabetes vom Typ I von Mein et al. (1998) einer weiteren Analyse unterzogen.

4.4.1 Datensatz von Risch (1990)

Die von Risch (1990) verwendeten IBD-Verteilungen sind in Tab. 4.5 aufgeführt. Nur 32 der 74 Paare sind bezüglich ihrer IBD-Verteilung voll informativ.

Mit der LOD-Score Transformation $T^2/(2\ln 10)$ ergibt sich für die voll informativen Paare und den klassischen Mittelwerttest T_m ein LOD-Score von 1.95 (Tab. 4.6). Wie zu erwarten war, ist der entsprechende LOD-Score des Tests mit der empirischen Varianz $T_{m,ev}$ mit 2.63 etwas höher. Der gewichtete Ansatz läßt beide klassischen Ansätze weit hinter sich zurück und zeigt einen LOD-Score von 3.09. Die empirischen p-Werte von 0.00094 bzw. 0.00079 zeigen an dieser Stelle allerdings kaum Unterschiede.

	Voll informative Familien		Alle Familien	
Teststatistik	LOD-Score	emp. p-Wert	LOD-Score	emp. p-Wert
MLS & TTS	2.20	–	2.79	–
T_m	1.95	–	1.60	–
$T_{m,ev}$	2.63	0.000942	2.77	0.000338
T_w	3.09	0.000793	3.25	0.000339

Tabelle 4.6: Ergebnisse von Maximum LOD Score (MLS), Holman's Triangle Test Statistic (TTS), Mittelwerttest ohne und mit empirischer Varianz (T_m, $T_{m,ev}$) im Vergleich zum gewichteten Mittelwerttest (T_w) anhand der Daten aus Tab. 4.5 . Empirische p-Werte wurden aus nach dem in 4.2.3 beschriebenen Verfahren mit $M = 1.000.000$ Replikaten ermittelt.

Die geschätzte IBD-Verteilung der vollständig informativen Paare beträgt $(0.06, 0.5, 0.44)$ und liegt somit innerhalb des possible-triangle (Holmans, 1993). Aus diesem Grund werden sowohl der MLS, als auch die Triangle-Test-Statistic (TTS) mit einem LOD-Score von 2.20 angegeben.

Werden die nicht-vollständig informativen Familien mit in die Berechnungen einbezogen, verringert sich der LOD-Score des klassischen Tests. Der Mittelwerttest mit empirischer Varianz verbessert sich leicht auf einen LOD-Score von 2.77, MLS und TTS auf einen Score von 2.79. Der LOD-Score des gewichteten Ansatzes steigt von 3.09 auf 3.20. Der Vergleich der empirischen p-Werte zeigt hier allerdings keinen Unterschied zwischen gewichtetem und ungewichtetem Ansatz. Dies mag auf die vergleichsweise geringe Fallzahl zurückzuführen sein (vgl. Abb. 4.2).

4.4.2 Datensatz von Mein et al. (1998)

Der Datensatz von Mein et al. (1998) beinhaltet den Genomscan von 356 erkrankten Geschwisterpaaren aus Großbritannien. Abbildung 1 ihrer Arbeit

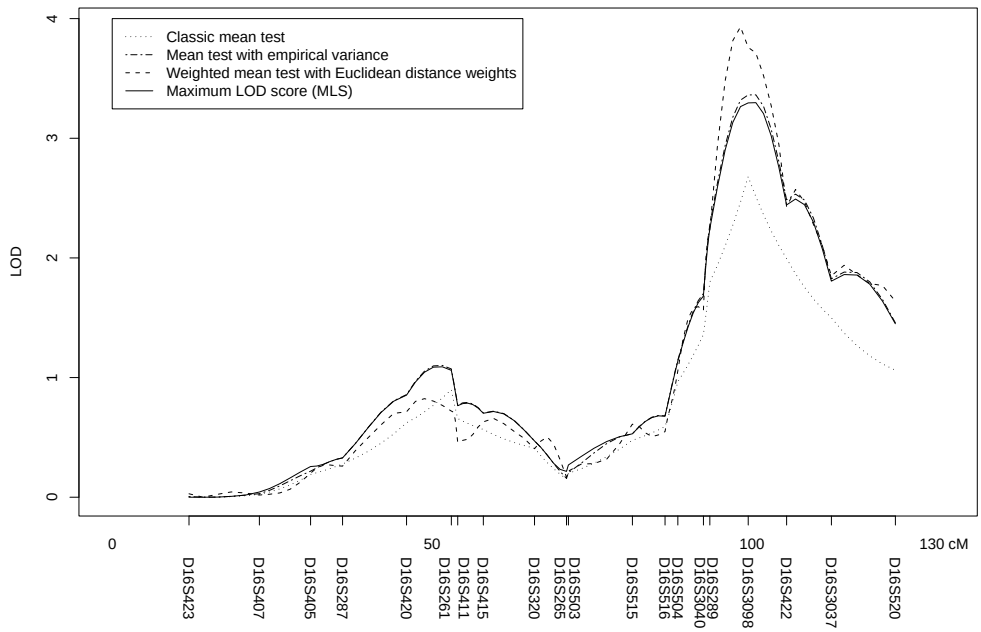


Abbildung 4.3: Mehrpunktanalyse von Chromosom 16, Daten von Mein et al. (1998). Der MLS wurde unter der Annahme von Dominanzvarianz berechnet, der klassische Mittelwerttest unter H_0, für den gewichteten Test wurde die Gewichtung nach euklidischem Abstand verwendet (Franke und Ziegler, 2005).

zeigt einen Kandidatengenort für Typ I Diabetes auf Chromosom 16q22-q24 (D16S515-D16S520), die Informativität des Scans lag durchweg zwischen 65% und 90%, am Maximum bei etwa 80% (vgl. Mein et al., 1998, Abb. 1a).

Die Autoren verwendeten zur Mehrpunktanalyse die Triangle Test Statistic (TTS) von Holmans (1993). Die Auswertung erfolgte in GENEHUNTER (Kruglyak und Lander, 1998) unter Berücksichtigung von Dominanzvarianz. Für diese Arbeit wurden die Daten von Mein et al. (1998) ein weiteres Mal analysiert. Zum einen wurden die Ergebnisse des TTS reproduziert, zum an-

	$T_{m,ev}$		T_w	
Position	LOD-Score	emp. p-Wert	LOD-Score	emp. p-Wert
93.4	2.277495	0.0006528	2.232573	0.0007789
94.6	2.627533	0.0003032	2.938877	0.0001680
95.8	2.938974	0.0001367	3.480040	0.0000460
97.0	3.176865	0.0000762	3.816359	0.0000212
98.2	3.318917	0.0000582	3.931289	0.0000166
99.4	3.362353	0.0000508	3.760043	0.0000265
100.6	3.363086	0.0000545	3.712628	0.0000297
101.8	3.267519	0.0000626	3.523601	0.0000429
103.0	3.075123	0.0000982	3.272137	0.0000735
104.2	2.804194	0.0001853	2.948014	0.0001557
105.4	2.486136	0.0004162	2.429615	0.0005076

Tabelle 4.7: Auszug der in Abb. 4.3 dargestellten Daten, hier als Vergleich zwischen Mittelwerttest ($T_{m,ev}$) und gewichtetem Mittelwerttest (T_w): gegeben sind für einige genetische Positionen in cM jeweils die LOD-Scores, berechnet aus $T^2/(2\ln 10)$, sowie die nach 4.2.3 ermittelten empirischen p-Werte ($M = 10.000.000$ Replikate).

deren die die hier vorgestellten Teststatistiken berechnet (Gleichungen 4.1 bis 4.3). Die sich ergebenden LOD-Score-Kurven sind in Abbildung Abb. 4.3 zusammengestellt. Der maximale LOD-Score des klassischen Ansatzes T_m liegt hier bei etwa 2.8, sowohl der Mittelwerttest mit empirischer Varianz als auch die TTS erreichen einen LOD von etwa 3.4. Der größte Score von etwa 3.9 wird vom gewichteten Mittelwerttest mit euklidischer Gewichtung erreicht. Das Ergebnis wird durch die nach Abschnitt 4.2.3 ermittelten empirischen p-Werte ($M = 10.000.000$) untermauert: die p-Werte des gewichteten Ansatzes sind signifikant kleiner als die des klassischen Mittelwerttests (Position 98.2, 0.000016 zu 0.000058). Insgesamt ist es bemerkenswert, daß die vier LOD-Score-Kurven sich in ihrer allgemeinen Form sehr ähnlich sind, die gewichtete Statistik dabei aber einen deutlich höheren Score erreicht.

Kapitel 5

Diskussion

Gewichtete Teststatistiken wurden in der genetischen Epidemiologie primär im Zusammenhang mit multiplen Geschwisterschaften diskutiert (Amos und Elston, 1989; Suarez und Hodge, 1979; Hodge, 1984). Suarez und Hodge (1979) zeigten, daß Kernfamilie mit s Nachkommen in etwa die Informativität von $(2s - 3 + 0.5^{s-1})/1.5$ unabhängigen Geschwisterpaaren besitzen. Somit entspricht, vom Standpunkt der Informativität, eine Geschwisterschaft der Größe s weder $s - 1$ unabhängigen, noch den $s(s - 1)/2$ möglichen, Geschwisterpaaren. Demzufolge schlug Hodge vor, Familien mit multiplen Geschwistern proportional zu ihrem Informationsgehalt zu gewichten. Blackwelder und Elston (1985) zeigten jenoch, daß multiple Geschwisterschaften einer Familie trotzdem asymptotisch als paarweise unabhängig angenommen werden können. Daher schlugen diese Autoren vor, erkrankte Geschwisterpaare grundsätzlich als unabhängig zu betrachten.

Auf quantitative Phänotypen lassen sich diese Resultate allerdings nicht anwenden. Für diese analysierten Wilson und Elston (1993) den Effekt mul-

tipler Geschwisterschaften und schlugen vor, Familien entsprechend ihres Informationsgehaltes, der auch hier von der Größe der betrachteten Familie abhängt, zu gewichten. Gewichtete Teststatistiken wurden in der Genetischen Epidemiologie ebenfalls zur optimalen Kombination verschiedener Teststatistiken (Sham und Purcell, 2001; Sham et al., 2002) oder verschiedener Studiendesigns vorgeschlagen (Gu et al., 1996; Ziegler et al., 1997).

Bei keinem der bisher diskutierten Ansätze spielt dabei die Markerinformativität eine Rolle: der Ansatz von Suarez und Hodge (1979) nutzte die Größe der Familie als Informationsmaß, Wilson und Elston (1993) verwendeten die Korrelation der quantitativen Phänotypen. Werden Teststatistiken oder verschiedene Studiendesigns miteinander kombiniert, kann als Informationsmaß die Varianz der Teststatistiken eingesetzt werden.

Obwohl also gewichtete Teststatistiken in der Genetischen Epidemiologie bereits in vielerlei Hinsicht untersucht wurden, wurde die Gewichtung von Familien nach Markerinformativität erst vor wenigen Jahren erstmalig betrachtet. Bisher wurde in der Literatur die Gewichtung einzelner Familien entsprechend des Informationsgehaltes der Marker nicht systematisch erarbeitet. So kritisierten noch Schork und Greenwood (2004) eine Verzerrung der klassischen Teststatistiken für modellfreie Kopplungsanalysen und zeigten in ihrer Arbeit auf, daß die Berücksichtigung nichtinformativer bzw. nur partiell informativer Familien zu einer erheblichen Verringerung der statistischen Macht der Kopplungsanalysen führen kann. Erstmalig angewendet wurde die Idee der Gewichtung nach Markerinformativität im Rahmen des Genetic Analysis Workshop 13 von Jacobs et al. (2003). Die Autoren gewichteten, unter Verwendung des Verfahrens von Haseman und Elston (1972) für

den quantitativen Phänotyp Blutdruck, Geschwisterschaften anhand des Informationsgehaltes der Marker. Die Autoren griffen dabei auf die klassischen asymptotischen Eigenschaften der ungewichteten Teststatistik von Haseman und Elston zurück. Sie zeigten allerdings nicht, daß die asymptotische Normalität auch dann gilt, wenn die Geschwisterschaften, wie durchgeführt, nach Markerinformativität gewichtet werden.

Ziel dieser Arbeit war daher zum einen, die von Jacobs et al. (2003) erstmals publizierte Idee zur Gewichtung von Geschwisterschaften nach Markerinformativität systematisch zu untersuchen. Zum anderen wurde, entsprechend dem von Schork und Greenwood aufgezeigten Problem, ein Ansatz für die Analyse von erkrankten Geschwisterpaaren entwickelt, der eine Gewichtung der einzelnen Geschwisterpaare nach ihrer jeweiligen Markerinformativität erlaubt.

Im ersten Teil dieser Arbeit wurde das Verfahren von Haseman und Elston (1972) um die Gewichtung nach Markerinformativität erweitert. Dabei wurde nachgewiesen, daß die von Jacobs et al. (2003) vorgeschlagene Modifikation der Teststatistik von Haseman und Elston asymptotisch keiner Standardnormalverteilung folgt. Daraus ergibt sich, daß die von Jacobs et al. (2003) berichteten Ergebnisse zu liberal waren. Es konnte außerdem gezeigt werden, daß jedes nicht-uniforme Gewichtungsschema zu einer Teststatistik führt, die, unter der Nullhypothese, nicht asymptotisch standardnormalverteilt ist. Wie gezeigt wurde, lassen sich p-Werte, unter Verwendung eines geeigneten Gewichtungsschemas, beliebig designen. Diese Problematik kann auf zweierlei Weisen angegangen werden: Einerseits besteht die Möglichkeit durch Permutation der Phänotypen empirische p-Werte zu bestimmen. Andererseits

könnte anstelle des auf der Fisher-Information basierenden Varianzschätzers, der robuste Varianzschätzer (vgl. Ziegler et al., 1998) verwendet werden. Die daraus resultierende Teststatistik folgt dann unter der Nullhypothese einer Standardnormalverteilung. Es zeigt sich allerdings, daß der Einsatz des robusten Varianzschätzers die statistische Macht des gewichteten Ansatzes, gegenüber dem ungewichteten Ansatz, nicht erhöht (Ergebnisse nicht präsentiert).

Für erkrankte Geschwisterpaare gelang im zweiten Teil der Arbeit der Nachweis, daß der gewichtete Mittelwertstest, bei der die Geschwisterpaare anhand der jeweiligen Markerinformativität gewichtet werden, in bestimmten Modellen eine größere statistische Macht besitzt als der ungewichtete Mittelwertstest. Insbesondere konnte auch die asymptotische Verteilung der gewichteten Mittelwertstatistik hergeleitet werden. Allerdings zeigte sich in Monte-Carlo Simulationsstudien auch, daß in vielen monogenen statistische Modellen der hier vorgeschlagene gewichtete Test dem ungewichteten nicht überlegen ist. Weitere Analysen zur Überprüfung, ob andere Gewichtungsschemata den beobachteten Unterschied in der Informativität zwischen Familien besser ausnutzen können, sind notwendig. Prinzipiell wäre eine numerische Maximierung der Teststatistik über die Gewichte möglich. Allerdings ließen sich dann keine Aussagen mehr über die asmyptotischen Eigenschaften der Teststatistik treffen. Entsprechend müßte die Ermittlung des p-Wertes des Tests über Permutationen oder Monte-Carlo Simulationen erfolgen. Dabei ist zu beachten, daß der in dieser Arbeit entwickelte neue Ansatz zur Monte-Carlo Simulation von p-Werten im gewichteten Mittelwertstest hierfür nicht geeignet wäre.

Die in dieser Arbeit vorgestellte neue Methode zur empirischen Bestimmung von p-Werten in modellfreien Kopplungsanalysen mit erkrankten Geschwisterpaaren simuliert, neue Geschwisterpaare anhand der berechneten Gewichte der beobachteten Paarungen. Für den zuvor genannten Ansatz der Maximierung der Teststatistik könnte in zukünftigen Studien untersucht werden, welche Restriktionen der Generierung eines neuen Datensatzes auferlegt werden müßten.

Die Gültigkeit der beschriebenen Simulationsmethode für empirische p-Werte wurde nur für Familien mit genau einem erkrankten Geschwisterpaar gezeigt. Zusätzlich wurde angenommen, daß beide Eltern genotypisiert sind, so daß eine Abhängigkeit von der Markerallelfrequenz weitestgehend ausgeschlossen wurde. Weitere Studien sind erforderlich, um die Eigenschaften des neuen Ansatzes zur empirischen Bestimmung von p-Werten sowie der gewichteten Mittelwertsteststatistik beispielsweise bei fehlenden elterlichen Genotypen, Vorliegen von Genotypisierungsfehlern bzw. misspezifizierten Allelfrequenzen zu analysieren.

Kapitel 6

Zusammenfassung

Obwohl in der genetisch-epidemiologischen Literatur die Gewichtung von Familien nach verschiedenen Informationsmaßen einen breiten Raum einnimmt, wurden bisher keine systematischen Untersuchungen zur Berücksichtigung der Markerinformativität in Kopplungs- oder Assoziationsanalysen durchgeführt. Schork und Greenwood (2004) machten auf dieses Problem aufmerksam. Mit dieser Arbeit wurde die aufgezeigte Lücke geschlossen. Es wurden systematisch modellfreie kopplungsanalytische Verfahren analysiert, in denen Kernfamilien mit zwei Kinden gemäß der beobachteten Markerinformativität gewichtet wurden.

Für quantitative Phänotypen wurde dabei gezeigt, daß die Verwendung des klassischen allgemeinen linearen Modells in Teststatistiken resultiert, die unter der Nullhypothese „keine Kopplung“ das nominelle Niveau nicht halten. Durch geschickte Wahl der Gewichtungsfunktion lassen sich vielmehr p-Werte beliebig designen. In einer Monte-Carlo Simulationsstudie konnte belegt werden, daß die von Jacobs et al. (2003) verwendete Gewichtung zu

liberalen Testergebnissen führt und die berichteten Kopplungsergebnisse zum Blutdruck nicht korrekt sind. Schließlich wurde gezeigt, daß die Gewichtung nach Markerinformativität beim Verfahren von Haseman und Elston keinen Gewinn an statistischer Güte bringt, wenn für die Bestimmung der Eigenschaften der Teststatistik die eingeführte Gewichtung adäquat berücksichtigt wird.

Für binäre Phänotypen hingegen konnte sehr wohl eine Erhöhung der statistischen Macht bei Gewichtung der erkrankten Geschwisterpaare nach Markerinformativität gezeigt werden. Für dieses Verfahren wurde desweiteren ein neues Monte-Carlo Simulationsverfahren zur empirischen Bestimmung von p-Werten hergeleitet.

Literaturverzeichnis

Amos, C. I. und Elston, R. C. Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet Epidemiol*, 6:349–360, 1989.

Amos, C. I., Elston, R. C., Wilson, A. F. und Bailey-Wilson, J. E. A more powerful robust sib-pair test of linkage for quantitative traits. *Genet Epidemiol*, 6:435–449, 1989.

Bailey-Wilson, J. E., Almasy, L., Andrade, M., Bailey, J., Bickeböller, H., Cordell, H. J., Daw, E. W., Goldin, L., Goode, E. L., Gray-McGuire, C., Hening, W., Jarvik, G., Maher, B. S., Mendell, N., Paterson, A. D., Rice, J., Satten, G., Suarez, B., Vieland, V., Wilcox, M., Zhang, H., Ziegler, A. und Maccluer, J. W. Genetic analysis workshop 14: microsatellite and single-nucleotide polymorphism marker loci for genome-wide scans. *BMC Genet*, 6:S1, 2005.

Bishop, D. T. und Williamson, J. A. The power of identity-by-state methods for linkage analysis. *Am J Hum Genet*, 46:254–265, 1990.

Blackwelder, W. C. und Elston, R. C. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol*, 2:85–97, 1985.

Büning, H. und Trenkler, G. *Nichtparametrische statistische Methoden*. Walter de Gruyter, 2. Auflage, 1994.

Collins, A. und Morton, N. E. Nonparametric tests for linkage with dependent sib pairs. *Hum Hered*, 45:311–318, 1995.

Day, N. E. und Simons, M. J. Disease susceptibility genes – their identification by multiple case family studies. *Tissue Antigens*, 8:109–119, 1976.

Dempfle, A. und Loesgen, S. Meta-analysis of linkage studies for complex diseases: an overview of methods and a simulation study. *Ann Hum Genet*, 68:69–83, 2004.

deVries, R. R. P., Fat, R. F. M., Lai, A., Nijenhuis, L. E. und Rood, J. J. Hla-linked genetic control of host response to mycobacterium leprae. *Lancet*, ii:1328–1330, 1976.

Drigalenko, E. How sib pairs reveal linkage. *Am J Hum Genet*, 63:1242–1245, 1998.

Elston, R. C., Buxbaum, S., Jacobs, K. B. und Olson, J. M. Haseman and elston revisited. *Genet Epidemiol*, 19:1–17, 2000.

Elston, R. C., Kringlen, E. und Namboodiri, K. K. Possible linkage relationships between certain blood groups and schizophrenia or other psychoses. *Behav Genet*, 3:101–106, 1973.

Ewens, W. J. und Grant, G. R. *Statistical Methods in Bioinformatics*. Springer Verlag, 1. Auflage, 2001.

Falconer, D. S. und Mackay, T. F. C. *Introduction to Quantitative Genetics*. Prentice Hall, 4. Auflage, 1996.

Franke, D., Kleensang, A. und Ziegler, A. Haseman-elston weighted by marker informativity. *BMC Genet*, 6:S50, 2005.

Franke, D., Kleensang, A. und Ziegler, A. Sibsim, quantitative phenotype simulation in extended pedigrees. *GMS Med Inform Biom Epidemiol*, 2: Doc04, 2006.

Franke, D. und Ziegler, A. Weighting affected sib pairs by marker informativity. *Am J Hum Genet*, 77:230–241, 2005.

Green, J. R. und Woodrow, J. C. Sibling method for detecting hla-linked genes in disease. *Tissue Antigens*, 9:31–35, 1977.

Gu, C., Todorov, A. und Rao, D. C. Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genet Epidemiol*, 13:513–533., 1996.

Haseman, J. K. und Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet*, 2:3–19, 1972.

Hodge, S. E. The information contained in multiple sibling pairs. *Genet Epidemiol*, 1:109–122, 1984.

Holmans, P. Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet*, 52:362–374, 1993.

Horvitz, D. G. und Thompson, D. J. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*, 47:663 – 685, 1952.

Idury, R. M. und Elston, R. C. A faster and more general hidden markov model algorithm for multipoint likelihood calculations. *Hum Hered*, 47:197 – 202, 1997.

Jacobs, K. B., Gray-McGuire, C., Cartier, K. C. und Elston, R. C. Genome-wide linkage scan for genes affecting longitudinal trends in systolic blood pressure. *BMC Genet*, 4:82, 2003.

Knapp, M., Seuchter, S. A. und Baur, M. P. Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests. *Hum Hered*, 44: 37 – 43, 1994.

Kruglyak, L. und Lander, E. S. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am J Hum Genet*, 57:439–454, 1995.

Kruglyak, L. und Lander, E. S. Faster multipoint linkage analysis using fourier transforms. *J Comput Biol*, 5:1–7, 1998.

Lehmann, E. L. und Romano, J. P. *Testing Statistical Hypotheses*. Springer, 3. Auflage, 2005.

Li, C. C. *Population Genetics*. University of Chicago Press, Chicago, IL, 1955.

Loesgen, S., Dempfle, A., Golla, A. und Bickeböller, H. Weighting schemes in pooled linkage analysis. *Genet Epidemiol*, 21:142–147, 2001.

Mein, C. A., Esposito, L., Dunn, M. G., Johnson, G. C., Timms, A. E., Goy, J. V., Smith, A. N., Sebag-Montefiore, L., Merriman, M. E., Wilson, A. J., Pritchard, L. E., Cucca, F., Barnett, A. H., Bain, S. C. und Todd, J. A. A

search for type 1 diabetes susceptibility genes in families from the united kingdom. *Nature Genet*, 19:297 – 300, 1998.

Ott, J. *Analysis of Human Genetic Linkage*. John Hopkins University Press, 3. Auflage, 1999.

R Development Core Team, . *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Risch, N. Linkage strategies for genetically complex traits. iii. the effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet*, 46:242–253, 1990.

S.A.G.E., . *Statistical Analysis for Genetic Epidemiology v4.6*, 2004. URL `http://darwin.cwru.edu/sage`.

Schaid, D. J., Olson, J. M., Gauderman, W. J. und Elston, R. C. Regression models for linkage: issues of traits, covariates, heterogeneity and interaction. *Hum Hered*, 55:86–96, 2003.

Schork, N. J. und Greenwood, T. A. Inherent bias toward the null hypothesis in conventional multipoint nonparametric linkage analysis. *Am J Hum Genet*, 74:306–316, 2004.

Sham, P. C. und Purcell, S. Equivalence between haseman-elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet*, 68:1527–1532, 2001.

Sham, P. C., Purcell, S., Cherny, S. S. und Abecasis, G. R. Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet*, 71:238–253, 2002.

Sham, P. C., Zhao, J. H. und Curtis, D. Optimal weighting scheme for affected sib-pair analysis of sibship-data. *Ann Hum Genet*, 61:61–69, 1997.

Shete, S., Tiwari, H. und Elston, R. C. On estimating the heterozygosity and polymorphism information content value. *Theor Popul Biol*, 57:265–271, 2000.

Strachan, T. und Read, A. P. *Human Molecular Genetics 3.* John Wiley & Sons: New York, 2004.

Strauch, K. *Kopplungsanalyse bei genetisch komplexen Erkrankungen mit genomischem Imprinting und Zwei-Genort-Krankheitsmodellen.* Dissertation, Rheinische Friedrich-Wilhelms-Universität, Sigmund-Freud-Straße 25, 53105 Bonn, 2002.

Suarez, B. K. und Hodge, S. E. A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. *Clin Genet*, 15: 126–136., 1979.

Suarez, B. K., Rice, J. und Reicht, T. The generalized sib pair ibd distribution: its use in the detection of linkage. *Ann Hum Genet*, 42:87–94, 1978.

Wan, Y., Cohen, J. und Guerra, R. A permutation test for the robust sib-pair linkage method. *Ann Hum Genet*, 61:79–87, 1997.

Whittemore, A. S. und Tu, I.-P. Simple, robust linkage tests for affected sibs. *Am J Hum Genet*, 62:1228–1242, 1998.

Wilson, A. F. und Elston, R. C. Statistical validity of the haseman-elston sib-pair test in small samples. *Genet Epidemiol*, 10:593–598, 1993.

Zhao, H., Merikangas, K. R. und Kidd, K. K. On a randomization procedure in linkage analysis. *Am J Hum Genet*, 65:1449–1456, 1999.

Zhao, J. H., Curtis, D. und Sham, P. C. Model-free analysis and permutation tests for allelic associations. *Hum Hered*, 50:133–139, 2000.

Ziegler, A. *Genetische Kartierung quantitativer Phänotypen: Eine Übersicht über modellfreie kopplungsanalytische Verfahren.* Number 84 in Medizinische Informatik, Biometrie und Epidemiologie. Urban & Vogel, Medien- und Medizin Verl., 1999.

Ziegler, A., Hebebrand, J. und Schäfer, H. A clinically orientated approach for combining discordant and concordant sib pairs [published erratum appeared in biom j 1997 39(7):880]. *Biom J*, 39:263–272, 1997.

Ziegler, A., Kastner, C. und Blettner, M. The generalised estimating equations: an annotated bibliography. *Biom J*, 40:115–139, 1998.

Ziegler, A. und König, I. R. *A statistical approach to genetic epidemiology.* Wiley-CVH Verlag GmbH & Co. KGaA, 2006.

Zinn-Justin, A., Ziegler, A. und Abel, L. Multipoint development of the weighted pairwise correlation (wpc) linkage method for pedigrees of arbi-

trary size and application to the analysis of breast cancer and alcoholism familial data. *Genet Epidemiol*, 21:40–52, 2001.

Danksagung

Bei Erstellung dieser Arbeit bin ich so einigen Menschen unendlich auf die Nerven gegangen. Bei diesen möchte ich mich entschuldigen und auch für ihre Unterstützung bedanken.

Zuallererst geht mein Dank natürlich an Prof. Dr. Andreas Ziegler der mir als Fachhochschulabsolent überhaupt erst die Möglichkeit gab eine Dissertation anzufertigen. Für die Überlassung des Themas bedanke ich mich ebenso, wie für die außerordentliche fachliche Betreuung. Desweiteren geht mein Dank auch an Frau Dr. Inke König, die als beratende Instanz immer für alle Fragen offen war.

Dank möchte ich Ph. D. Robert Elston und Ph. D. James Malley aussprechen, die mir Aufenthalte an ihren jeweiligen Einrichtungen in den USA ermöglichten.

Nicht zuletzt geht mein Dank auch an all die anderen Menschen die mit mir die letzten Jahre verbracht und mich ertragen haben, Arbeitskollegen, Freunde und Familie.

Diese Arbeit wurde sowohl durch die Deutsche Forschungsgemeinschaft (ZI 597/12-1) als auch das Bundesministerium für Bildung und Forschung (PGE-S26T11) finanziell unterstützt.

Lebenslauf

Persönliche Daten

Name	Daniel Franke
Anschrift	Ratzeburger Allee 88, 23562 Lübeck
Geburtsdatum	24.10.1976 in Waldshut-Tiengen
Familienstand	ledig
Staatsangehörigkeit	deutsch

Schul- und Hochschulausbildung

1983 – 1993	Hebelschule Laufenburg (Grundschule) Hochrhein-Gymnasium, Waldshut-Tiengen Robert-Schuhmann Realschule Waldshut-Tiengen
1993	Mittlere Reife
1993 – 1996	Friedrich Weinbrenner Gewerbeschule, Freiburg i. Br.
1996	Abschluß der Lehre zum Vermessungstechniker
1997 – 1998	Gewerbeschulen Bad Säckingen
1998	Fachhochschulreife
1998 – 2002	Studium der Mathematik an der FH Darmstadt
2002	Abschluß des Mathematikstudiums mit Diplom (FH)
seit 2002	Wissenschaftlicher Mitarbeiter am Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck

Publikationen

Franke, D., Kastner, C. und Ziegler, A. Generalized estimating equations for association structures: Familial correlations of lipid profiles. *Stat in Med*, 23:855–857, 2004.

Franke, D., Kleensang, A. und Ziegler, A. Haseman-elston weighted by marker informativity. *BMC Genet*, 6:S50, 2005a.

Franke, D., Kleensang, A. und Ziegler, A. Sibsim, quantitative phenotype simulation in extended pedigrees. *GMS Med Inform Biom Epidemiol*, 2: Doc04, 2006.

Franke, D., Philippi, A., Tores, F., Hager, J., Ziegler, A. und König, I. R. On confidence intervals for genotype relative risks and attributable risks from case parent trio design for candidate-gene studies. *Hum Hered*, 60:81–88, 2005b. Erratum in *Hum Hered*, 60:180, 2005.

Franke, D. und Ziegler, A. Weighting affected sib pairs by marker informativity. *Am J Hum Genet*, 77:230–241, 2005.

Kleensang, A., Franke, D., Koenig, I. R. und Ziegler, A. Genomwide haplotype sharing analysis for alcohol dependance based on quantitative traits and mantel statistic. *BMC Genet*, 6:S75, 2005.

Simon, M., Franke, D., Ludwig, M., Köster, G., Aliashkevich, A. F., Oldenburg, J., Ince, A., Ziegler, A. und Schramm, J. Association of a polymorphism of the alk1 gene (acvrl1) with sporadic arteriovenous malformations of the cns. *J Neurosurg*, 2005. in press.