Der Liebermeister-Test mit SAS®

Bernd Paul Jäger
Inst. f. Biometrie u. Med. Informatik
Ernst-Moritz-Arndt-Universität
Walther-Rathenau-Str. 48
17487 Greifswald
bjaeger@biometrie.uni-greifswald.de

Evgenija Klassen
Inst. f. Biometrie u. Med. Informatik
Ernst-Moritz-Arndt-Universität
Walther-Rathenau-Str. 48
17487 Greifswald
evgenijaklassen@gmx.de

Karl-Ernst Biebler
Inst. f. Biometrie u. Med. Informatik
Ernst-Moritz-Arndt-Universität
Walther-Rathenau-Str. 48
17487 Greifswald
biebler@biometrie.uni-greifswald.de

Paul Eberhard Rudolph
Forschungsinstitut für die Biologie
landwirtschaftlicher Nutztiere (FBN)
Wilhelm-Stahl-Allee 2
18196 Dummerstorf
pe.rudolph@fbn-dummerstorf.de

**Herrn Prof. Dr. med. habil. Günther Kraatz,
Direktor der Klinik und Poliklinik für Innere Medizin A
der Ernst-Moritz-Arndt-Universität Greifswald,
dem akademischen Lehrer und Freund,
zum 65. Geburtstag gewidmet**

Zusammenfassung

Für die zweiseitige statistische Testentscheidung beim Vergleich zweier Binomialwahrscheinlichkeiten kann aufgrund der vorliegenden Simulationsergebnisse im Falle kleiner Stichprobenumfänge anstelle des üblichen exakten Tests von Fisher der Liebermeister-Test empfohlen werden. Er hält das Signifikanzniveau ein, ist aber weniger konservativ. In Hinsicht auf den Fehler 2. Art kann, abhängig von der betrachteten Situation, der exakte Test von Fisher oder der Liebermeister-Test vorteilhafter sein.
Der Liebermeister-Test ist mit SAS® unter Verwendung von PROC FREQ realisierbar, wenn die Daten wie im Vergleich von Tabelle 1 und Tabelle 2 ersichtlich variiert werden.

Schlüsselwörter: exakter Test von Fisher, Liebermeister-Test, Konfidenzintervalle für die Odds ratio, Konfidenzintervall-basierte Tests

1 Einleitung

Beobachtet werden zwei binomialverteilte Zufallsgrößen mit den Parametern (p_1, n_1) bzw. (p_2, n_2).
Zum Vergleich der Wahrscheinlichkeiten p_1 und p_2 im Falle zweier unabhängiger Stichproben findet oft der χ^2- Test Anwendung. Bei kleinen Stichprobenumfängen ist

dieser nicht zu empfehlen, weil die Prüfgröße durch die asymptotische χ^2- Verteilung nur unzureichend beschrieben wird.
Für genau diese Testsituation stehen der exakte Fisher-Test, der Liebermeister-Test und ein auf exakten Konfidenzintervallen der Wahrscheinlichkeiten p_1 und p_2 basierender Test zur Verfügung.
In medizinischen Publikationen, insbesondere auf dem Gebiet der Epidemiologie, werden häufig aus Interpretationsgründen nicht die Wahrscheinlichkeiten p_1 und p_2, sondern daraus abgeleitete Maße, wie die Risikodifferenz $RD = p_1 - p_2$, das relative Risiko $RR = p_1 / p_2$ oder das Chancenverhältnis (Odds ratio) $OR = (p_1 / (1 - p_1)) / (p_2 / (1 – p_2))$ betrachtet.
Aus statistischer Sicht sind diese Begriffe kein neuer Zugang zum Testproblem, denn die entsprechenden Hypothesen H_0: $p_1 = p_2$, H_0: $RD = 0 = p_1 - p_2$, H_0: $RR = 1 = p_1 / p_2$ und H_0: $OR = 1$ sind äquivalent. Möchte man aber konsequent bei den neu eingeführten Begriffen bleiben, dann steht der leichten Interpretation die schwierigere (abgesehen von der Risikodifferenz) statistische Behandlung des Testproblems gegenüber. Für den Chancenquotienten OR wird das vorgeführt.
Als Methoden zur Abschätzung von Fehlern 1. Art α und 2. Art β wurden durchgängig Simulationsverfahren mit SAS® unter Verwendung der Module BASE [4] und STAT [5] genutzt.

2 Die untersuchten Methoden

2.1 Liebermeister-Test und exakter Test von Fisher

Im Jahr 1877 entwickelte Carl Liebermeister einen kombinatorischen statistischen Test zum Vergleich von Erfolgswahrscheinlichkeiten unterschiedlicher Therapien für kleine Stichproben, den Liebermeister-Test. Das war mehr als ein halbes Jahrhundert bevor der heute für die gleiche Testsituation zur Anwendung kommende exakte Fisher-Test ersonnen wurde. Im Laufe der Zeit ist der Liebermeister-Test in Vergessenheit geraten, der exakte Test von Fisher ist heute in den meisten Statistik-Programmsystemen verfügbar. Dabei sind beide Tests sehr ähnlich.

Carl Liebermeister kam am 22.Februar 1833 in Ronsdorf als erstes von neun Kindern in einer wohlhabenden Kaufmannsfamilie zur Welt. Er studierte Medizin in Bonn, Würzburg und Greifswald. Seine ersten Berufsjahre verbrachte er an der Greifswalder Klinik für Innere Medizin, die in jenen Jahren (1855-1860) von Felix Niemeyer geführt wurde. Als Niemeyer 1862 eine Berufung nach Tübingen erhielt, folgte Liebermeister seinem akademischen Lehrer. 1865 nahm er den Ruf als Professor für Innere Medizin in Basel an. Er kehrte 1871 nach Tübingen zurück, um die Nachfolge des Lehrstuhls von Felix Niemeyer anzutreten. Er starb an einem Nierentumor am 24. November 1901 (nach [7]).

Bei beiden Tests geht man von einer Vier-Felder-Tafel aus. Beim exakten Test von Fisher werden darin die beobachteten Anzahlen niedergelegt. Sie hat das folgende Aussehen:

Tabelle 1: Vier-Felder-Tafel für kategoriale Daten im Zweistichprobenfall

	Erfolg der Behandlung	kein Erfolg	gesamt
Stichprobe 1	m	n_1 - m	n_1
Stichprobe 2	k - m	n_2 - (k - m)	n_2
Gesamt	k	$n_1 + n_2$ - k	$n_1 + n_2$

Schätzungen für die Erfolgswahrscheinlichkeiten p_1 und p_2 ergeben sich aus den Formeln $p_1 = m / n_1$ und $p_2 = (k - m) / n_2$. Getestet werden:

H_0: $p_1 = p_2$ gegen H_A: $p_1 > p_2$ beim linksseitigen Test,
H_A: $p_1 < p_2$ beim rechtsseitigen Test und
H_A: $p_1 \neq p_2$ beim zweiseitigen Test.

Die Nullhypothese H_0 wird abgelehnt, falls $P < \alpha$ gilt. P steht dabei für den P-Wert des entsprechenden Tests.

P-Wert für den Fisher-Test:

$$P_F = \sum_{r \geq m} \frac{\binom{n_1}{r}\binom{n_2}{k-r}}{\binom{n_1+n_2}{k}}$$

P-Wert für den Liebermeister-Test:

$$P_L = \sum_{r \geq m+1} \frac{\binom{n_1+1}{r}\binom{n_2+1}{k+1-r}}{\binom{n_1+n_2+2}{k+1}}$$

Formal kann man sich die Prüfgröße des Liebermeister-Tests aus der des exakten Fisher-Tests entstanden denken, wenn man in der Stichprobe 1 den Stichprobenumfang um ein Erfolgsereignis und in der Stichprobe 2 den Umfang um ein Misserfolgsereignis erhöht. Man erhält damit die folgende Vier-Felder-Tafel (siehe Tab. 2), auf die der exakte Test von Fisher angewandt wird.
Der Liebermeister-Test ist also der exakte Test von Fisher bezüglich der leicht variierten Beobachtungszahlen. Auf diese Weise kann der Liebermeister-Test mit Hilfe des in SAS® vorhandenen exakten Fisher-Tests ausgeführt werden.

Tabelle 2: Variation der Vier-Felder-Tafel nach Liebermeister

	Erfolg der Behandlung	kein Erfolg der Behandlung	gesamt
Stichprobe 1	$m + 1$	$n_1 - m$	$n_1 + 1$
Stichprobe 2	$k - m$	$n_2 - (k - m) + 1$	$n_2 + 1$
Gesamt	$k + 1$	$n_1 + n_2 - k + 1$	$n_1 + n_2 + 2$

Aus der Literatur (s. [3]) ist bekannt, dass beide Tests unterschiedliches Verhalten zeigen. Der exakte Test von Fisher ist im linksseitigen Fall konservativ, d.h. er tendiert dazu, die Nullhypothese länger aufrecht zu erhalten. Der Liebermeister-Test ist quasi-exakt, d.h. sein tatsächlicher Fehler 1. Art bleibt nahe an der Signifikanzschwelle. Durch seine Vorgehensweise, ähnlich einer „Kontinuitätskorrektur" die Erfolge in Stichprobe 1 und die Misserfolge in Stichprobe 2 zu vergrößern, verschiebt er im Vergleich zum Fisher-Test die Entscheidung von der Nullhypothese in Richtung Alternativhypothese. Dass der Liebermeister-Test das α – Risiko sowohl im linksseitigen als auch im zweiseitigen Fall immer noch einhält, wird in den nachfolgend vorgestellten Simulationsexperimenten bestätigt.

2.2 Konfidenzintervall-basierter Test, beruhend auf dem Binomialansatz

Der exakte Test von Fisher und der Liebermeister-Test werden mit einem Konfidenzintervall-basierten Test verglichen. Man geht von der Vier-Felder-Tafel für den exakten Test von Fisher aus (Tab. 1).
Aus den Tafeldaten schätzt man mit der Maximum-Likelihood-Methode die Erfolgswahrscheinlichkeiten für jede der beiden Stichproben:

$$\hat{p}_1 = \frac{m}{n_1} \qquad \text{bzw.} \qquad \hat{p}_2 = \frac{k-m}{n_2}.$$

Bezüglich der beiden Stichproben werden exakte Konfidenzintervalle zum Niveau α für die Erfolgswahrscheinlichkeiten p_1 und p_2 berechnet. Jeweils mit der Wahrscheinlichkeit α wird also der unterliegende Parameter nicht überdeckt und mit $1-\alpha$ wird er überdeckt. Mit diesen Konfidenzintervallen KI_1 für p_1 und KI_2 für p_2 werden die folgenden Tests für H_0 begründet:
Es wird zum einen überprüft, ob die Schätzung $\hat{p}_1$ im Konfidenzintervall KI_2 liegt, dann kommt $\hat{p}_1$ als möglicher Erfolgsparameter der zweiten Stichprobe in Frage. Überdeckt das Konfidenzintervall KI_2 die Schätzung $\hat{p}_1$ aber nicht, dann kann $\hat{p}_1$ nicht als Erfolgsparameter der zweiten Stichprobe gelten und die Hypothese H_0: $p_1 = p_2$ wird abgelehnt.
Die gleichen Argumente für die Schätzung $\hat{p}_2$ und das Konfidenzintervall KI_1 begründen einen zweiten Test.
Sind die Stichprobenumfänge ungleich, so liefert die größere der beiden Stichproben das kleinere Konfidenzintervall und der Test ist „entscheidungsfreudiger", die kleinere Stichprobe mit dem entsprechend größeren Konfidenzintervall ist konservativer.

Mit Hilfe von links- bzw. rechtsseitigen Konfidenzintervallen lassen sich analog links- bzw. rechtsseitige Tests begründen.
Die Berechnung der exakten Konfidenzintervalle wird mit Hilfe des SAS® -Makros CI-BINOM realisiert, das der Literatur (vgl. [1]) entnommen ist.

2.3 Tests, basierend auf den Konfidenzintervallen für den Chancenquotienten (Odds ratio)

In der angloamerikanischen medizinischen Fachliteratur hat sich der Begriff „Odds ratio“ weit verbreitet. Er wird aus dem Begriff „Chance“ (englisch: odds) abgeleitet. Mit Wahrscheinlichkeit wird der Anteil der für das Experiment günstigen Fälle bezogen auf die möglichen Fälle relativiert. Bei Chance, einem Begriff aus der Wettpraxis, wird das Verhältnis von den für das Experiment günstigen Fällen zu den für das Experiment ungünstigen Fällen beschrieben. Beide Begriffe sind ineinander überführbar, aus der Wahrscheinlichkeit p erhält man die Chance Ch = p / (1- p) und aus der Chance Ch die Wahrscheinlichkeit p = Ch / (1+ Ch).

Im Zweistichprobenfall ergeben sich in Analogie zum relativen Risiko $RR = \frac{p_1}{p_2}$, das mit Hilfe der assoziierten Chancen ermittelte Chancenverhältnis oder Odds ratio OR,

$$OR = \frac{Ch_1}{Ch_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

Bei kleinen Wahrscheinlichkeiten p_1 und p_2 fallen die Begriffe relatives Risiko RR und Odds ratio OR wegen 1- $p_1 \approx 1$ und 1- $p_2 \approx 1$ näherungsweise zusammen. Der Vorteil beider Relativzahlen RR und OR gegenüber der alleinigen Betrachtung von p_1 und p_2 liegt in der „gängigeren“ Interpretation.
Dieser Vorteil der einfachen Interpretation wird aber durch die Schwierigkeiten bei der Berechnung der Konfidenzgrenzen für RR und OR verspielt.
Weil man die Verteilung der Schätzungen $\hat{p}_1$ und $\hat{p}_2$ für die Wahrscheinlichkeiten p_1 und p_2 kennt (erwartungstreue Maximum-Likelihood-Schätzungen mit Minimalvarianz), kann man sie asymptotisch durch eine Normalverteilung annähern,

$$\hat{p}_i \sim N\left(p_i, \frac{p_i(1-p_i)}{n_i^2}\right),\ i = 1,2\,.$$

Die Verteilungen der Quotienten

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2},\ \hat{Ch}_1 = \frac{\hat{p}_1}{1-\hat{p}_1},\ \hat{Ch}_2 = \frac{\hat{p}_2}{1-\hat{p}_2} \text{ oder } \hat{OR} = \frac{\hat{Ch}_1}{\hat{Ch}_2}$$

sind zunächst unbekannt und können erst nach weiteren Transformationen und dann auch nur näherungsweise bestimmt werden. Im Weiteren werden drei Näherungsverfahren zur Berechnung der Konfidenzgrenzen für die Odds ratio angegeben, nämlich diejenigen nach Woolf, nach Miettinen sowie ein so genanntes exaktes Verfahren. Die Ver-

fahren werden kommentarlos angegeben. Eine Bewertung wird im Anschluss an das Simulationsexperiment erfolgen.

2.3.1 Konfidenzintervall für die Odds ratio nach Woolf

Der Umweg über den Logarithmus der Odds ratio $\log(\hat{OR})$, approximiert durch ein Taylorpolynom vom Grade 1, führt zur asymptotischen Varianz

$$V(\log(\hat{OR})) \approx \frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_2} + \frac{1}{n_2 - k_2}$$

und damit zu einem angenäherten Konfidenzintervall zum Niveau α für die transformierte Zufallsgröße $\log(\hat{OR})$,

$$\log(\hat{OR}) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_2} + \frac{1}{n_2 - k_2}},$$

wobei $u_{1-\frac{\alpha}{2}}$ das entsprechende Quantil der Normalverteilung sowie $k_1 = m$ und $k_2 = k - m$ die Anzahl der Erfolge in Stichprobe 1 und 2 sind. Das Konfidenzintervall $\left(\hat{OR}_u ; \hat{OR}_o\right)$ für $\hat{OR}$ wird durch Rücktransformation aus dem des $\log(\hat{OR})$ erhalten:

$$\hat{OR}_{o/u} = \exp\left(\log(\hat{OR}) \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{k_1} + \frac{1}{n_1 - k_1} + \frac{1}{k_2} + \frac{1}{n_2 - k_2}} \right).$$

In der Literatur finden sich zahlreiche Hinweise, dass das so berechnete Konfidenzintervall für kleine Stichprobenumfänge ungeeignet ist. In [3] findet man als Faustformeln für die Anwendung die Forderung $n > 20$. Zum anderen sollte man 2 als Häufigkeit jeder der Zellen der Vierfeldertafel bei $n < 100$ nicht unterschreiten.

2.3.2 Konfidenzintervall für die Odds ratio nach Miettinen

Das Konfidenzintervall $\left(\hat{OR}_u ; \hat{OR}_o\right)$ für die Odds ratio nach Miettinen bestimmt man aus

$$\hat{OR}_u = \min\left(\hat{OR}^{1-\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\chi^2}}}, \hat{OR}^{1+\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\chi^2}}} \right) \text{ und } \hat{OR}_o = \max\left(\hat{OR}^{1-\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\chi^2}}}, \hat{OR}^{1+\frac{u_{1-\frac{\alpha}{2}}}{\sqrt{\chi^2}}} \right),$$

wobei $u_{1-\frac{\alpha}{2}}$ das $\left(1-\frac{\alpha}{2}\right)$- Quantil der Standardnormalverteilung N(0,1) und χ^2 die Prüfgröße des χ^2-Homogenitätstestes der zugehörigen Vier-Felder-Tafel sind (siehe auch [2]).

2.3.3 Exakte Methode zur Bestimmung der Konfidenzgrenzen der Odds ratio

Neben den Näherungsmethoden gibt es auch eine so genannte exakte Methode zur Bestimmung der Konfidenzgrenzen der Odds ratio, die auf kombinatorischen Überlegungen ähnlich denen des exakten Fisher-Tests beruht.
Für $n_u = \max(k_1 + k_2 + n_1 - n, 0)$ und $n_o = \min(k_1 + k_2, n_1)$, werden das $x = OR_u$ gesucht, das die Gleichung

$$\frac{\sum_{i=n_u}^{k_1} \binom{k_1 + k_2}{i} \cdot \binom{n - k_1 - k_2}{n_1 - i} x^i}{\sum_{i=n_u}^{n_o} \binom{k_1 + k_2}{i} \cdot \binom{n - k_1 - k_2}{n_1 - i} x^i} = \frac{\alpha}{2}$$

erfüllt, und das $x = OR_o$, das die Gleichung

$$\frac{\sum_{i=k_1}^{n_o} \binom{k_1 + k_2}{i} \cdot \binom{n - k_1 - k_2}{n_1 - i} x^i}{\sum_{i=n_u}^{n_o} \binom{k_1 + k_2}{i} \cdot \binom{n - k_1 - k_2}{n_1 - i} x^i} = \frac{\alpha}{2} \text{ erfüllt.}$$

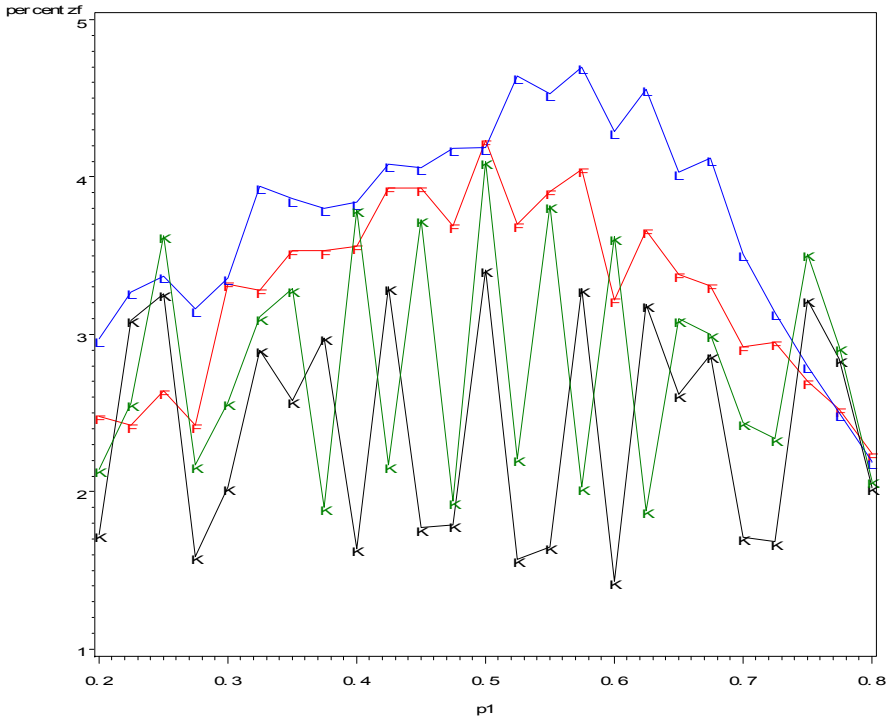
Man beachte, dass nur die Summationsgrenzen des Zählers in beiden Gleichungen verschieden sind!
In SAS® (vgl. [3], [4]) ist die aufwändige Berechnung der exakten oberen und unteren Konfidenzgrenzen der Odds ratio OR in der PROC FREQ möglich, wenn die Option „exact“ angegeben wird. Standardmäßig werden aber die Näherungen nach Woolf angegeben.

3 Simulationsexperimente mit SAS®

Die Vier-Felder-Tafel ist die Grundlage der Tests. Mit Hilfe eines Zufallsexperiments werden 10 000 Tafeln erzeugt, die jeweils aus zwei unabhängigen Realisierungen binomialverteilter Zufallsgrößen mit den Parametern n_1, p_1 bzw. n_2, p_2 mit vorgegebenen Erfolgswahrscheinlichkeiten p_1 und p_2 entstehen. Dafür werden jeweils die Testgrößen berechnet und die statistischen Entscheidungen getroffen. So kann zum einen die Wahrscheinlichkeitsverteilung der Prüfgröße empirisch nachgebildet und zum anderen geprüft werden, wie genau die relative Häufigkeit der Ablehnungen der Nullhypothese mit dem vorher festgelegten Wert $\alpha = 0.05$ übereinstimmt. Für die Konfidenzintervalle geht man ebenfalls von den 10 000 Vier-Felder-Tafeln aus. Es werden die Punkt- und die Konfidenzschätzungen ($\alpha = 0.05$) der Odds ratio nach Woolf, Miettinen und nach der exakten Methode bestimmt.

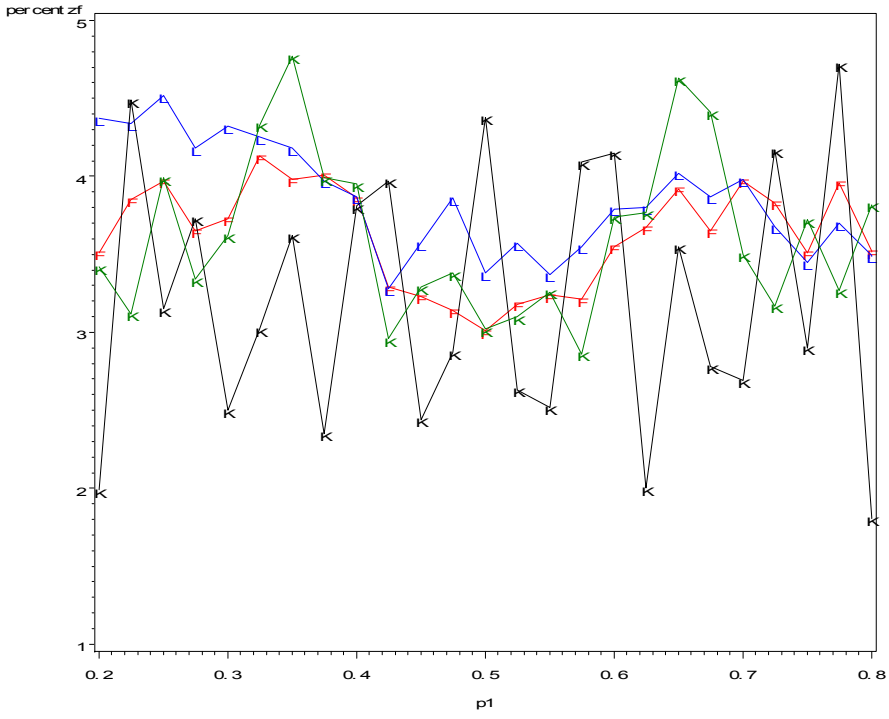
3.1 Vergleich der Fehler 1. Art des Liebermeister-Tests, des exakten Tests von Fisher und des Konfidenzintervall-basierten Tests für die Erfolgswahrscheinlichkeiten

Zunächst vergleicht man den Fehler 1. Art der Tests, indem man mit gleichen Erfolgswahrscheinlichkeiten ($p_1 = p_2$) von 0.2 bis 0.8 mit einer Schrittweite von 0.1 Stichproben erzeugt. Als Stichprobenumfänge wurden n_1=15 und n_2=20 (man spricht dabei noch von balancierten Stichproben), n_1=15 und n_2=50 sowie n_1=30 und n_2=80 gewählt.



n_1=15, n_2=20

Legende
Abszisse: p1
Ordinate: Fehler 1. Art in %

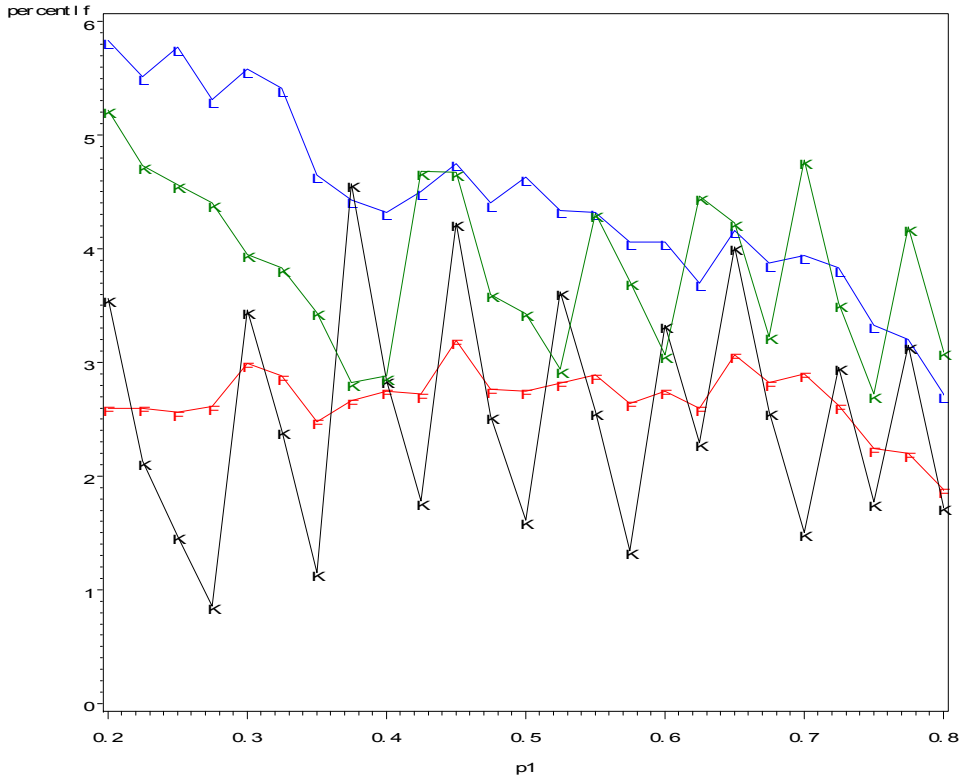


n1=30, n2=80

Legende
Abszisse: p_1
Ordinate: Fehler 1. Art in %

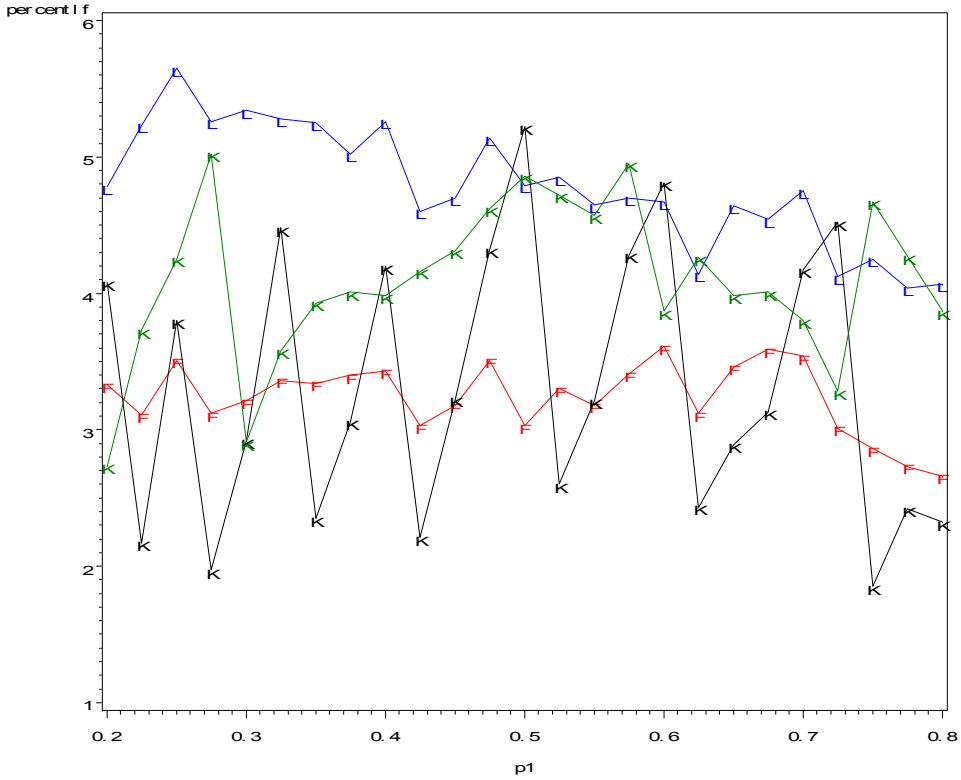
Abbildung 1: Vergleich der Fehler 1. Art des exakten Tests von **F**isher, des **L**iebermeister-Tests und der beiden Konfidenzintervall-basierten Tests der Erfolgswahrscheinlichkeiten (vgl. Abschnitt 2.2) im zweiseitigen Fall in Abhängigkeit von $p = p_1 = p_2$ für verschiedene Stichprobenumfänge

Die experimentell ermittelten prozentualen Fehler 1. Art der Konfidenzintervall-basierten Tests liegen im balancierten Falle fast immer unter den Werten der anderen Tests (Abbildung 1). Dies geht mit wachsender Unbalanciertheit verloren.
Sowohl bei linksseitiger, als auch bei zweiseitiger Alternative liegen die experimentell ermittelten prozentualen Fehler 1. Art des Liebermeister-Tests stets über den experimentell ermittelten prozentualen Fehlern 1. Art des exakten Tests von Fisher. Im rechtsseitigen Fall ist es genau umgekehrt. Man kann dies damit erklären, dass beim Liebermeister-Test ein Erfolg zur Stichprobe 1 hinzugezählt wird, der die Erfolgswahrscheinlichkeit erhöht, und gleichzeitig ein Misserfolg zur 2. Stichprobe addiert wird, der die Erfolgswahrscheinlichkeit der Stichprobe 2 verringert. Dadurch steigt die Chance des Liebermeister-Tests gegenüber dem Fisher-Test, die Nullhypothese abzulehnen.
Bei kleinen Stichprobenumfängen n_1 und n_2 überschreitet der Liebermeister-Test beim linksseitigen Fall die Signifikanzschwelle $\alpha = 0.05$ (s. Abb. 2, $n_1 = 15$, $n_2 = 20$, beispielsweise für $p < 0.3$). Im rechtsseitigen Fall ist der Liebermeister-Test extrem konservativ. Das tatsächliche Niveau liegt zwischen 0.01 und 0.02 und damit weit unter dem angestrebten Niveau $\alpha = 0.05$. Den Liebermeister-Test sollte man beim zweiseitigen und beim linksseitigen Test gegenüber dem exakten Fisher-Test favorisieren.



n_1=15, n_2=20

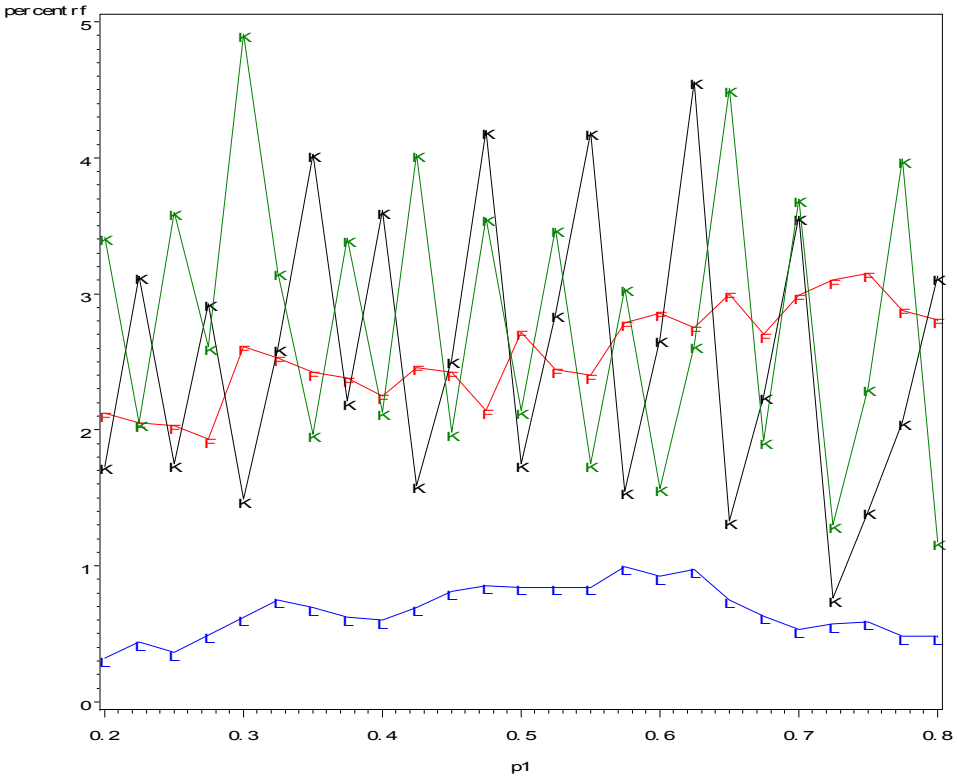
Legende
Abszisse: p_1
Ordinate: Fehler 1. Art in %



n_1=30, n_2=80

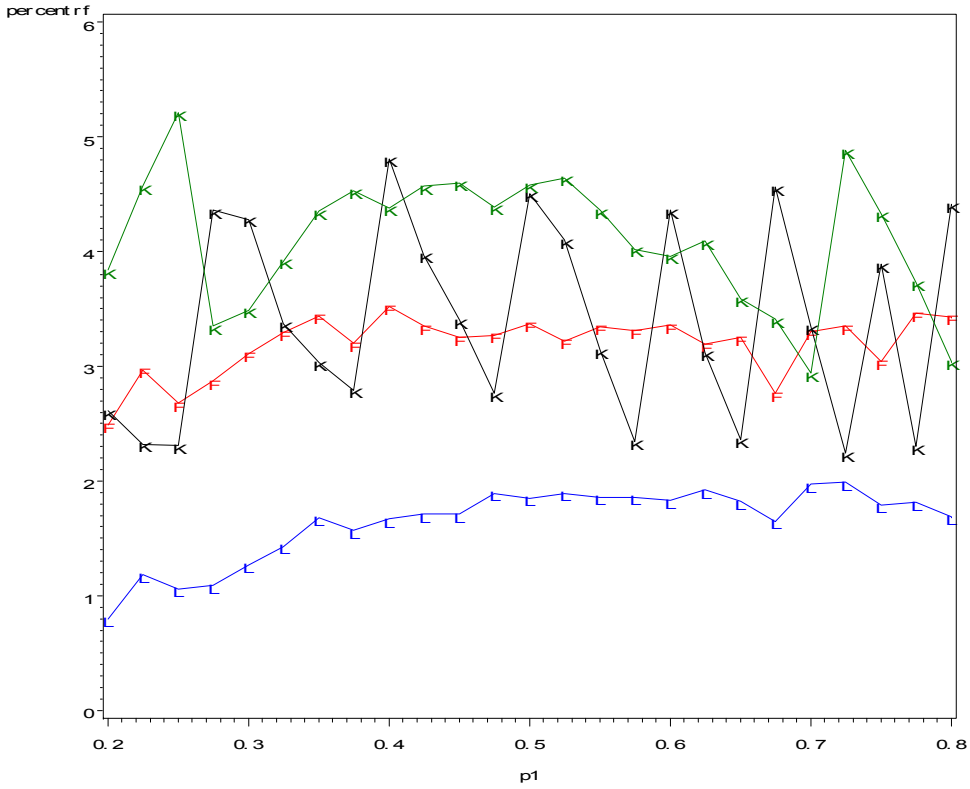
Legende
Abszisse: p_1
Ordinate: Fehler
1. Art in %

Abbildung 2: Vergleich der Fehler 1. Art des **L**iebermeister-Tests, des exakten Test von **F**isher und der beiden Konfidenzintervall-basierten Tests der Erfolgswahrscheinlichkeiten (vgl. Abschnitt 2.2) im linksseitigen Fall in Abhängigkeit von $p = p_1 = p_2$ für verschiedene Stichprobenumfänge



n_1=15, n_2=20

Legende
Abszisse: p_1
Ordinate: Fehler
1. Art in %



n_1=30, n_2=80

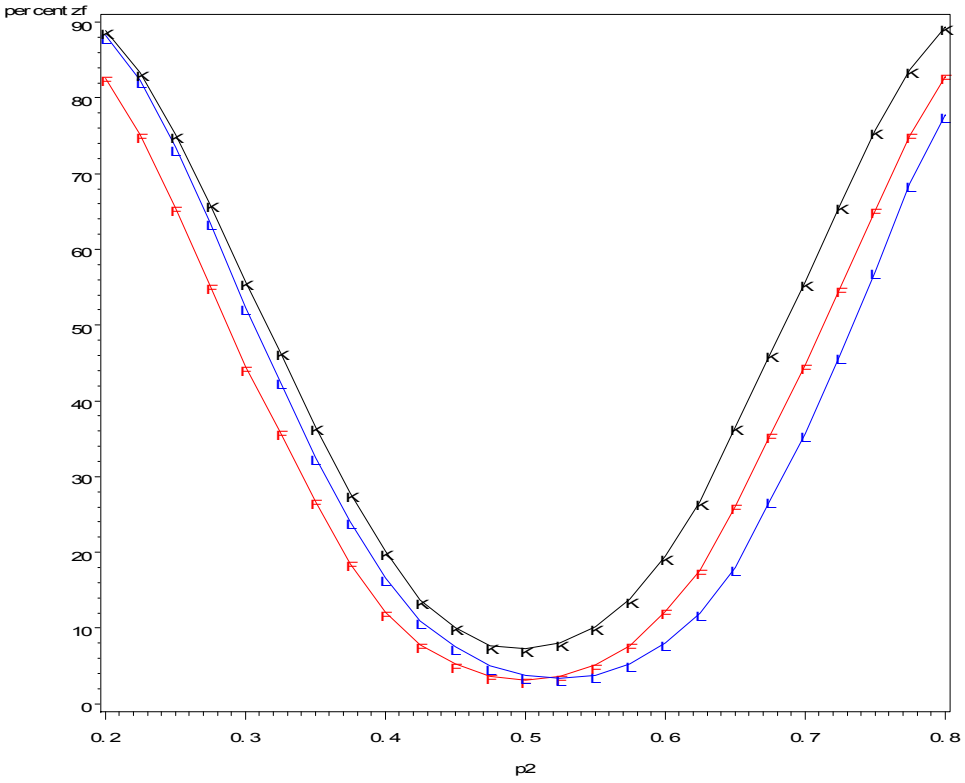
Legende
Abszisse: p_1
Ordinate: Fehler 1. Art in %

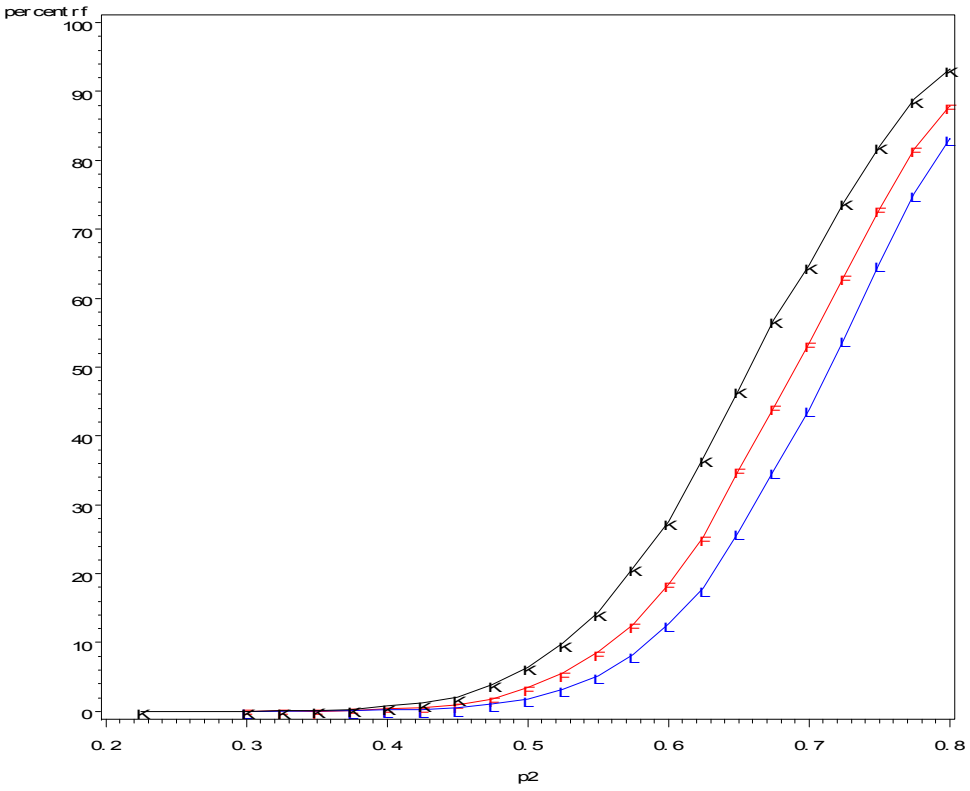
Abbildung 3: Vergleich der Fehler 1. Art des **L**iebermeister-Tests, des exakten Test von **F**isher und der beiden Konfidenzintervall-basierten Tests der Erfolgswahrscheinlichkeiten (vgl. Abschnitt 2.2) im rechtsseitigen Fall in Abhängigkeit von $p = p_1 = p_2$ für verschiedene Stichprobenumfänge

3.2 Vergleich der Teststärke des Liebermeister-Tests, des exakten Tests von Fisher und des Konfidenzintervall-basierten Tests

In diesem Abschnitt wird die Teststärke von Tests untersucht. Dazu betrachtet man zwei binomialverteilte Stichproben vom Umfang n_1=30 und n_2=80, fixiert p_2 bei 0.5 (weil bei dieser Parameterwahl die Varianz der Binomialverteilung am größten ist) und lässt p_1 in den Grenzen zwischen 0.1 bis 0.9 mit der Schrittweite 0.025 variieren. Auf die entstehenden Vier-Felder-Tafeln werden der exakte Test von Fisher, der Liebermeister-Test und ein auf dem exakten Konfidenzintervall beruhender Test (Konfidenzintervall aus der größeren Stichprobe) angewendet.
In der Abbildung 4 ist der in 10 000 Simulationen ermittelte Fehler 2. Art als Prozentwert angegeben. Man sieht: Im zweiseitigen Fall ist der Fehler 2. Art des Konfidenzintervall-basierten Tests am höchsten. Im zweiseitigen Fall macht der Liebermeister-Test für kleine p größere Fehler 2. Art als der exakte Test, für große p ist es genau umgekehrt. Im linksseitigen Fall ist der Liebermeister-Test der stärkste, gefolgt von exaktem Test und konfidenzbasiertem Test. Im rechtsseitigen Fall ist der β – Fehler des Konfidenzintervall-basierten Tests ebenfalls wieder am höchsten, gefolgt vom Liebermeister-Test und dem exakten Test von Fisher.

Für andere Testsituationen (n_1=15, n_2=20 und n_1=15, n_2=50) ergaben sich analoge Resultate.





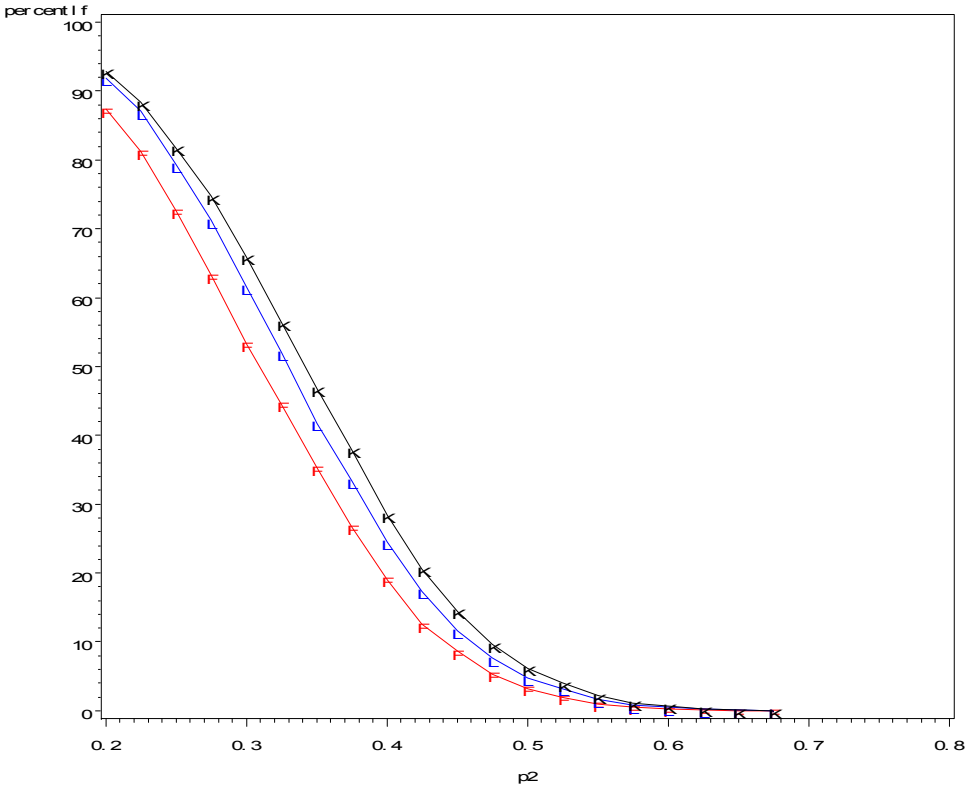


Abbildung 4: Vergleich der Fehler 2.Art (in %) des exakten Tests von **F**isher, des **L**iebermeistertests und des Konfidenzintervall-basierten Tests für die Stichprobenumfänge n_1 =30 und n_2 = 80 im zwei-, links- und rechtsseitigen Fall (von oben nach unten)

3.3 Vergleich der näherungsweise kalkulierten Konfidenzintervalle für die Odds ratio

Dem Anwender stehen zur Berechnung der Konfidenzgrenzen der Odds ratio mehrere Möglichkeiten zur Verfügung. Welche soll er wählen?
Erinnert sei daran, dass ein Konfidenzintervall den Parameter mit mindestens der Wahrscheinlichkeit $1-\alpha$ überdeckt oder mit höchstens der Wahrscheinlichkeit α nicht überdeckt. In einem Simulationsprogramm werden zwei binomialverteilte Stichproben mit den Stichprobenumfängen n_1 und n_2 und vorgegebenem OR gezogen, die Konfidenzintervalle nach den verschiedenen oben beschriebenen Methoden gebildet und nachgesehen, ob diese den vorgegebenen Parameter OR überdecken. Dieses Experiment wird zehntausendmal wiederholt. Für jedes Berechnungsverfahren zählt man, wie viele Nichtüberdeckungen eintraten. Ein „ordnungsgemäßes“ Konfidenzintervall sollte in nicht mehr als 500 von 10 000 Fällen (für $\alpha = 0.05$) den vorgegebenen Parameter OR nicht überdecken.
Ein Problem bei diesem Rechenexperiment bestehen darin, dass zu einem OR mehrere Paare (p_1, p_2) existieren, die zum gleichen OR führen. Mit diesem Paar werden aber die beiden Stichproben simuliert. Die größte Varianz bringt man in das Simulationsverfahren, wenn p_1 = 0.5 gewählt wird. In diesem Falle variieren vermutlich auch das ge-

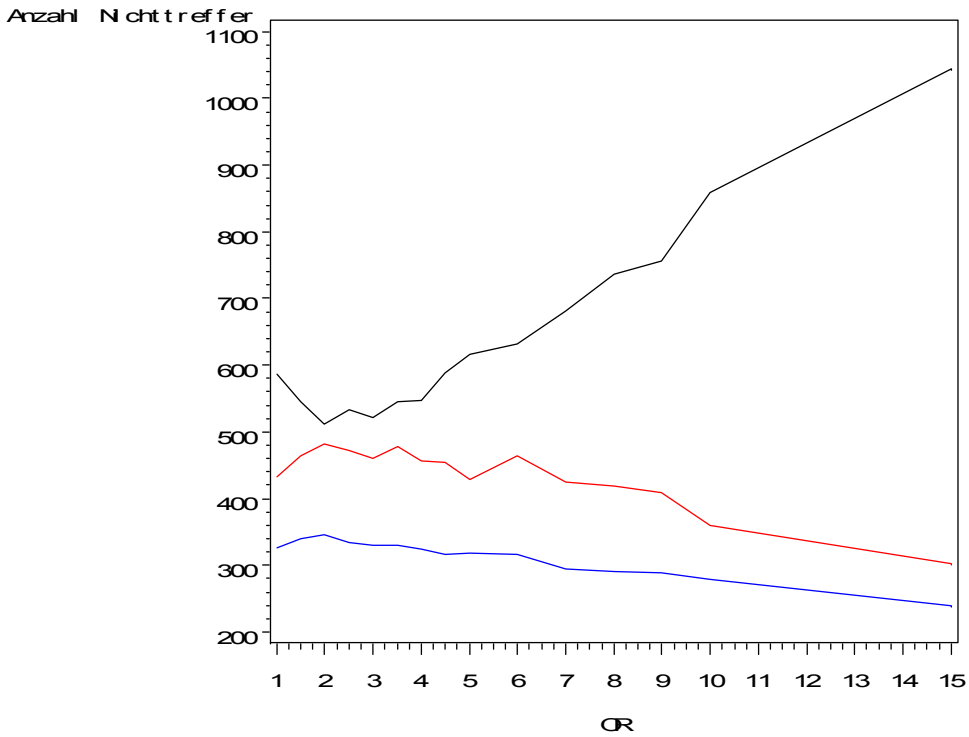
schätzte OR und damit die Breiten der Konfidenzintervalle am meisten. Dieser ungünstigste Fall wird simuliert.
Die erste Stichprobe wird aus einer Binomialverteilung $B(n_1, 0.5)$ gezogen. Ein vorgegebenes OR führt dann wegen

$$OR = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} \Leftrightarrow p_2 = \frac{p_1}{OR(1-p_1)+p_1}$$

zu einem p_2 und einer zweiten binomialverteilten Stichprobe

$$B\left(n_2, \frac{0.5}{OR \cdot 0.5 + 0.5}\right).$$

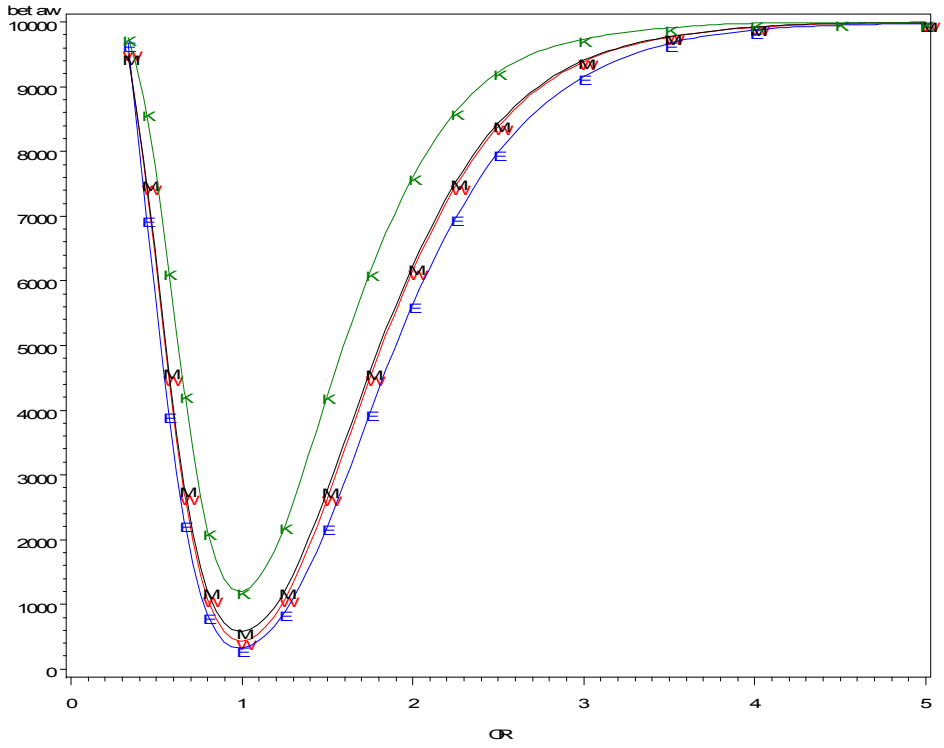
Auf diese Weise wurden 10 000 zufällige Vier-Felder-Tafeln erzeugt und daraus die Konfidenzintervalle nach Woolf, Miettinen sowie dem exakten Verfahren berechnet. Jeweils war zu ermitteln, ob das berechnete Konfidenzintervall den vorgegebenen Parameterwert für OR überdeckt. Vorgegeben war ein Konfidenzniveau von 0.05.



Legende
Abszisse: OR
Ordinate: Anzahl der Nichtüberdeckungen von OR durch das Konfidenzintervall

Abbildung 5: Anzahl von Nichtüberdeckungen der vorgegebenen Odds ratio OR durch das exakte Konfidenzintervall (untere Linie), durch das Konfidenzintervall von Woolf (mittlere Linie) und durch das Konfidenzintervall nach Miettinen (obere Linie) bei durchgeführten 10 000 Simulationen

Die Ergebnisse der Simulation sind in der Abbildung 5 dargestellt. Das exakte Konfidenzintervall weist für alle betrachteten Werte von OR die geringste Anzahl von Nichtüberdeckungen unter den drei Typen von Konfidenzintervallen auf. Das Konfidenzintervall nach Woolf hält für alle untersuchten Werte von OR das Konfidenzniveau ein und besitzt erst mit größerem OR eine höhere Überdeckungswahrscheinlichkeit als 0.95. Das Konfidenzintervall der Odds ratio nach Miettinen hält das vorgegebene Konfidenzniveau auch bei kleinem OR nur näherungsweise ein und ist für größer werdende OR vollkommen inakzeptabel.
Das Verfahren von Woolf kann nach diesen Simulationsexperimenten als „bestes“ im Sinne der Einhaltung des Konfidenzniveaus betrachtet werden. Es wird also zu Recht von SAS® bei der Berechnung der Konfidenzgrenzen für OR favorisiert.
Nimmt man diese Konfidenzintervalle als Basis für einen statistischen Test, so muss der Nullhypothese entsprechend nach Überdeckung von $OR = 1$ gefragt werden.
Die simulierten Fehler 2. Art β für statistische Tests, beruhend auf den Konfidenzintervallen der Odds ratio nach der exakten Methode, nach der Methode von Woolf und nach der Methode von Miettinen, sind in Abbildung 6 dargestellt als Anzahlen von Nichtüberdeckung von $OR = 1$ bei 10 000 durchgeführten Versuchen.
Zusätzlich wurde der Konfidenzintervall-basierte Test, beruhend auf einem exakten Konfidenzintervall für den Binomialparameter der größeren Stichprobe, wie oben beschrieben durchgeführt.



Legende
Abszisse: Odds ratio OR
Ordinate: Anzahl der Fehlentscheidungen bei 10 000 Simulationen

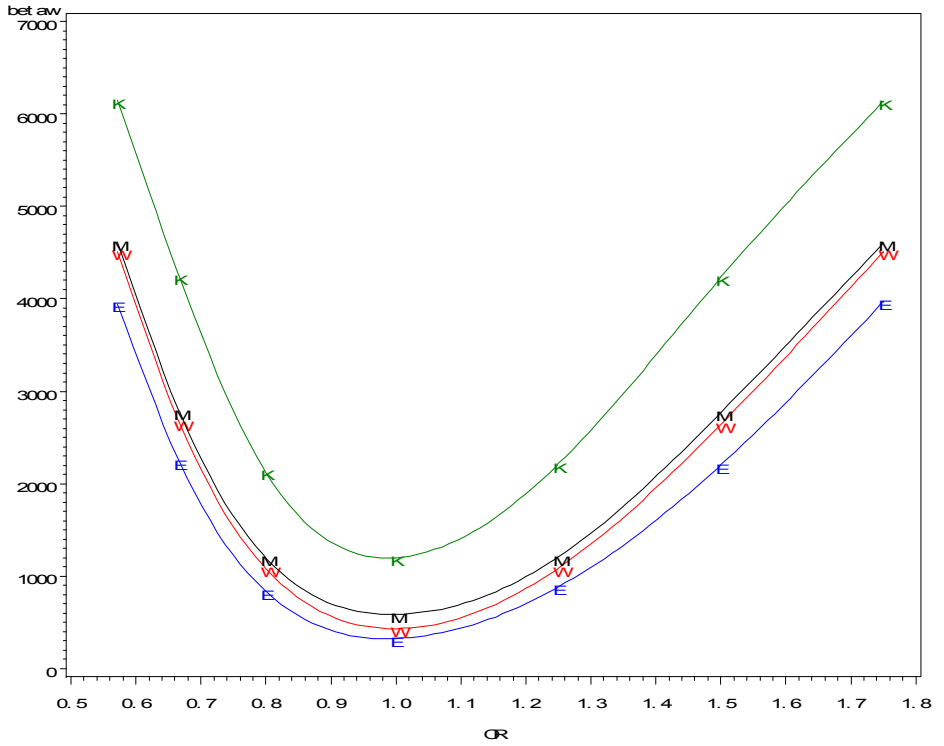


Abbildung 6: Fehler 2. Art β für statistische Tests in Abhängigkeit von OR, beruhend auf den Konfidenzintervallen der Odds ratio berechnet nach:
E der **e**xakten Methode, W der Methode nach **W**oolf, M der Methode nach **M**iettinen sowie K dem Test abgeleitet aus dem exakten **K**onfidenzintervalle für p
Unteres Bild: Vergrößerter Ausschnitt in der Nähe von OR = 1

Der auf dem exakten Konfidenzintervall der Binomialverteilung beruhende Test schneidet bezüglich des Fehlers β am schlechtesten ab. Ihm folgen die Konfidenzintervall-basierten Tests entsprechend den Verfahren von Miettinen und dann von Woolf. Wird das Konfidenzintervall für OR exakt berechnet, so besitzt der assoziierte Test den kleinsten Fehler 2. Art.

Literatur

[1] Daly, L. (1992). Simple SAS® macros for the calculation of exact binomial and Poisson confidence limits. Computers in Biology and Medicine 22, 5, pp. 351-361

[2] Lachin, J. M. (2000). Biostatistical methods, the assessment of relative risks. John Wiley & Sons, Inc., New York

[3] Rasch, D.; Herrendörfer, G.; Bock, J.; Victor, N.; Guiard, V. (1998). Verfahrensbibliothek: Versuchsplanung und –auswertung, R.Oldenbourgh Verlag München Wien. S.175-178

[4] SAS® Institute Inc. (2006). *Base SAS® 9.1.3 Procedures Guide, Second Edition, Volumes 1, 2, 3, and 4*. Cary, NC: SAS® Institute Inc.

[5] SAS® Institute Inc. (2004). *SAS/STAT 9.1 User's Guide*. Cary, NC: SAS® Institute Inc.

[6] Seneta, E.; Phipps, M. C. (2001). On the comparison of two observed frequencies. Biometrical Journal 43, 1, pp. 23-43

[7] Seneta, E.; Seif, F. J.; Liebermeister, H.; Dietz, K. (2004). Journal of Medical Biography 12, pp. 215-221

Anhang: SAS® -Programm

```
/* Programm vergleicht in einem Simulationsexperiment die
   Testergebnisse des exakten Fisher-Tests, des Liebermeister-Tests
   und der Tests, basierend auf exakten Konfidenzintervallen. */
%MACRO Liebermeister(umf1,umf2,p1,p2);
data fisher;
keep n1 n2 p1 p2 k i erfolg gruppe;
/* Festgelegt werden zwei Stichprobenumfänge und
   die beiden Erfolgswahrscheinlichkeiten p1 und p2. */
n1=&umf1;/* Stichprobenumfang 1. Stichprobe */
n2=&umf2;
p1=&p1;/* Binomialparameter 1.Stichprobe */
p2=&p2;
do k=1 to 10000;/* Simulationsumfang */
do i=1 to n1;
if UNIFORM(-1)<p1 then erfolg=1; else erfolg=0;
gruppe=1;
output;
end;/* 1. Stichprobe festgelegt durch Zufallsexperiment */
do i=1 to n2;
if UNIFORM(-1)<p2 then erfolg=1; else erfolg=0;
gruppe=2;
output;
end;/* 2. Stichprobe festgelegt durch Zufallsexperiment */
end;/* Datei work.fisher enthält 100000*(n1+n2) Datensätze! */
run;

/* proc means wird aufgerufen, um n1 n2 p1 p2 in Ausgabedatei
   means zu schreiben und an flag anzuhängen. */
proc means data=fisher noprint;
var n1 n2 p1 p2;
output out=means;
run;
data means;
set means;
keep n1 n2 p1 p2;
where _stat_="MAX"; run;

proc freq data=fisher noprint;
by k;
tables gruppe*erfolg/nocol nopercent exact;
output out=hilfef fisher;
/* Ergebn. des exakten Tests in Ausgabedatei hilfef umgeleitet. */
```

```
run;

data hilfef;
set hilfef;
/* Umbenennung der Systemvariablen in
   lf - linksseitige Wahrscheinlichkeit des exakten Fishertests,
   rf - rechtsseitige Wahrscheinlichkeit des exakten Fishertests,
   zf - zweiseitige Wahrscheinlichkeit des exakten Fishertests */
keep lf rf zf;
lf=XPL_FISH;
rf=XPR_FISH;
zf=XP2_FISH;
label lf='Fisher linkss.'
      rf='Fisher rechtss.'
      zf='Fisher zweis.';
run;

/* Die folgenden zwei data-Steps hängen bei abc_g1
   zusätzlich einen Erfolg an die Gruppe 1 und bei abc_g2
   einen Misserfolg an die Gruppe 2 an
   (für jedes k, d.h. jede Simulation). */
data abc_g1;
do k=1 to 10000;
erfolg=1;
gruppe=1;
output;
end;
run;
data abc_g2;
do k=1 to 10000;
erfolg=0;
gruppe=2;
output;
end;
run;


/* Zusammenfügen und Sortieren */
data liebermeister;
set fisher abc_g1 abc_g2;
run;
proc sort;
by k;
run;
```

```
/* Exakter Fisher-Test für die erweiterten Dateien
   ist der Liebermeistertest */
proc freq noprint;
by k;
tables gruppe*erfolg/nocol nopercent exact;
output out=hilfel fisher;
run;
data hilfel;
set hilfel;
keep ll rl zl;
/* Umbenennung der Systemvariablen */
ll=XPL_FISH;
rl=XPR_FISH;
zl=XP2_FISH;
label ll='Lieberm. linkss.'
      rl='Lieberm. rechtss.'
      zl='Lieberm. zweis.';
run;


/* Fisher- und Liebermeistertest in einer Datei zusammen */
data hilfe;
merge hilfef hilfel;
/* Kontrolle, wie die Tests entscheiden (für Power der Tests) */
if lf=. then Flag_lf=.;
if (lf>=0 and lf<=0.05) then Flag_lf=1;
if  lf>0.05 then Flag_lf=0;
/* Testentscheidung Fischer links */

if ll=. then Flag_ll=.;
if (ll>=0 and ll<0.05) then Flag_ll=1;
if  ll>0.05 then Flag_ll=0;
/* Testentscheidung Liebermeister links */

if rf=. then Flag_rf=.;
if (rf>=0 and rf<0.05) then Flag_rf=1;
if  rf>0.05 then Flag_rf=0;
/* Testentscheidung Fischer rechts */

if rl=. then Flag_rl=.;
if (rl>=0 and rl<0.05) then Flag_rl=1;
if  rl>0.05 then Flag_rl=0;
```

```
/* Testentscheidung Liebermeister rechts */

if zf=. then Flag_zf=.;
if (zf>=0 and zf<0.05) then Flag_zf=1;
if  zf>0.05 then Flag_zf=0;
/* Testentscheidung Fischer zweiseitig */

if zl=. then Flag_zl=.;
if (zl>=0 and zl<0.05) then Flag_zl=1;
if  zl>0.05 then Flag_zl=0;
/* Testentscheidung Liebermeister zweiseitig */

/* Bedeutung der Variablen:
   Flag_ll=1 - linksseitiger Liebermeistertest signifikant ,
   Flag_rl=1 - rechtsseitiger Liebermeistertest signifikant,
   Flag_zl=1 - zweiseitiger Liebermeistertest signifikant */
run;
/* Erfolge bzw. Misserfolge in beiden Gruppen
   werden für jede Simulation k zusammengezählt */
proc freq data=fisher noprint;
by k;
tables erfolg*gruppe / out=zwischen;
run;


data e1;
set zwischen;
keep k e1;
where gruppe=1 and erfolg=1;
e1=count;
run;

data e2;
set zwischen;
keep k e2;
where gruppe=2 and erfolg=1;
e2=count;
run;

/* Datei alles enthält Erfolge und Misserfolge Gruppe 1 u. 2. */
data alles;
merge e1 e2;
by k;
if e1=. then e1=0;
```

```
if e2=. then e2=0;
run;

/* Test basierend auf den Konfidenzintervallen */
/* Berechnung und graphische Darstellung des exakten
   Konfidenzgrenzen für den Parameter p der Binomialverteilung.
   Eingabe des Stichprobenumfangs N im Programmteil nötig.
   Programm nutzt das Makro CIBINOM.
INPUT PARAMETERS:
CL - Confidence level CL (should be 0.95)
N - Total sample size
R - Number with characteristic

OUTPUT PARAMETERS:
P - Observed proportion (R/N)
PL - Lower confidence limit
PU - Upper confidence limit

USAGE NOTES:
Missing values for the output parameters are generated if
(i) N is equal to zero, or if
(ii) R is less than zero or greater than N or if
(iii) CL is not between 0.0 and 1.0.
*/
%MACRO CIBINOM (CL,N,R,P,PU,PO);
IF ((&N) EQ 0) OR (NOT(0 LE (&R) LE (&N))) OR (NOT(0 LT (&CL) LT 1))
THEN DO;
&P=.;
&PU=.;
&PO=.;
END;

ELSE DO;
&P=(&R)/(&N);
IF (&R) EQ 0 THEN DO;
&PU=0;
&PO=1 - 10**(LOG10((1-(&CL))/2)/(&N));
END;
ELSE IF (&R) EQ (&N) THEN DO;
&PU=10**(LOG10((1-(&CL))/2)/(&N));
&PO=1;
END;
ELSE DO;
&PU=1 - BETAINV(((1+(&CL))/2),((&N)+1-(&R)),(&R));
```

```
&PO=BETAINV(((1+(&CL))/2),((&R)+1),((&N)-(&R)));
END;
END;
%MEND;

DATA KI1;
set alles;
alpha=0.05;
CL=1-alpha;
nn1=&umf1;
%CIBINOM (CL,nn1,e1,q1,puz1,poz1);
output;
run;

DATA KI2;
set alles;
alpha=0.05;
CL=1-alpha;
nn2=&umf2;
%CIBINOM (CL,nn2,e2,q2,puz2,poz2);
output;
run;

DATA KI3;
set alles;
alpha=0.1;
CL=1-alpha;
nn1=&umf1;
%CIBINOM (CL,nn1,e1,q1,pul1,pol1);
output;
run;

DATA KI4;
set alles;
alpha=0.1;
CL=1-alpha;
nn2=&umf2;
%CIBINOM (CL,nn2,e2,q2,pul2,pol2);
output;
run;

data hilfeKI;
merge KI1 KI2 KI3 KI4;
/* Kontrolle, wie der Test entscheidet (für Power der Tests) */
```

```
if puz1<q2 and q2<poz1 then flag_kz1=0;else flag_kz1=1;
if puz2<q1 and q1<poz2 then flag_kz2=0;else flag_kz2=1;
/* zweiseitig KI */
if pul1<q2 then flag_kr1=0;else flag_kr1=1;
if pul2<q1 then flag_kr2=0;else flag_kr2=1;
/* rechtsseitig */
if pol1>q2 then flag_kl1=0;else flag_kl1=1;
if pol2>q1 then flag_kl2=0;else flag_kl2=1;
/* linksseitig */
run;

/* Fishertest-Ergebnisse und Liebermeistertest-Ergebnisse in HILFE,
   KI-Test-Ergebnisse in hilfeKI, ab jetzt zusammen in HILFE */
data hilfe;
merge hilfe hilfeKI;
run;

/* Häufigkeit der Flags wird je Tests in Datei ausgegeben. */
proc freq data=hilfe noprint;
tables Flag_lf /outcum out=flaglf;
run;
data flaglf;
set flaglf;
keep percentlf;
percentlf=percent;
where flag_lf=1;
run;

proc freq data=hilfe noprint;
tables Flag_rf /outcum out=flagrf;
run;
data flagrf;
set flagrf;
keep percentrf;
percentrf=percent;
where flag_rf=1;
run;

proc freq data=hilfe noprint;
tables Flag_zf /outcum out=flagzf;
run;
data flagzf;
set flagzf;
keep percentzf;
```

```
percentzf=percent;
where flag_zf=1;
run;
/* Fisher fertig */

proc freq data=hilfe noprint;
tables Flag_ll /outcum out=flagll ;
run;
data flagll;
set flagll;
keep percentll;
percentll=percent;
where flag_ll=1;
run;

proc freq data=hilfe noprint;
tables Flag_rl /outcum out=flagrl ;
run;
data flagrl;
set flagrl;
keep percentrl;
percentrl=percent;
where flag_rl=1;
run;

proc freq data=hilfe noprint;
tables Flag_zl /outcum out=flagzl ;
run;
data flagzl;
set flagzl;
keep percentzl;
percentzl=percent;
where flag_zl=1;
run;
/* Liebermeister fertig */

proc freq data=hilfe noprint;
tables flag_kz1 /outcum out=flagk1;
run;
data flagk1;
set flagk1;
keep percentk1;
percentk1=percent;
where flag_kz1=1;
```

```
run;

proc freq data=hilfe noprint;
tables flag_kz2 /outcum out=flagk2;
run;
data flagk2;
set flagk2;
keep percentk2;
percentk2=percent;
where flag_kz2=1;
run;

proc freq data=hilfe noprint;
tables flag_kr1 /outcum out=flagkr1;
run;
data flagkr1;
set flagkr1;
keep percentkr1;
percentkr1=percent;
where flag_kr1=1;
run;

proc freq data=hilfe noprint;
tables flag_kr2 /outcum out=flagkr2;
run;
data flagkr2;
set flagkr2;
keep percentkr2;
percentkr2=percent;
where flag_kr2=1;
run;

proc freq data=hilfe noprint;
tables flag_kl1 /outcum out=flagkl1;
run;
data flagkl1;
set flagkl1;
keep percentkl1;
percentkl1=percent;
where flag_kl1=1;
run;

proc freq data=hilfe noprint;
tables flag_kl2 /outcum out=flagkl2;
```

```
run;
data flagkl2;
set flagkl2;
keep percentkl2;
percentkl2=percent;
where flag_kl2=1;
run;
/* Konfidenzintervall-basierter Test fertig */

/* Zusammenfügen der Dateien */
data flag;
merge means flaglf flagrf flagzf
flagll flagrl flagzl
flagk1 flagk2 flagkr1 flagkr2 flagkl1 flagkl2;
run;

/* Anfügen der flag-Datei an work.ergebnis innerhalb des Macro */
proc append base=ergebnis data=flag force;
run;
%MEND;

proc datasets;
delete ergebnis;
run;
quit;

%Liebermeister(15,20,0.5,0.2);

title1 'Vergleich des exakten Tests von Fisher,';
title2 'des Liebermeistertests und der Tests';
title3 'basierend auf den Konfidenzintervallen';

proc print data=ergebnis noobs;
var n1 n2 p2 percentlf percentrf percentzf
percentll percentrl percentzl
percentk1 percentk2 percentkl1 percentkl2 percentkr1 percentkr2;
run;
```