



ERNST-MORITZ-ARNDT-UNIVERSITÄT GREIFSWALD
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT
Institut für Mathematik und Informatik

Vergleich von Unabhängigkeitstests für stetige Größen

Wissenschaftliche Arbeit zur Erlangung
des akademischen Grades
„Diplom-Mathematiker“

vorgelegt von Marcus Vollmer

geboren am 27. April 1985
in Neubrandenburg

Universitäts- und Hansestadt Greifswald, 9. September 2010

Erstgutachter/Betreuer: Prof. Christoph Bandt

Zweitgutachter: Prof. Michael Schürmann

Danksagung

Ich möchte mich an dieser Stelle bei all denen bedanken, die mich während meines Studium begleitet und unterstützt haben. Liebe Freunde, vielen Dank für die vielen Spieleabende, den sportlichen Aktivitäten, dem gemeinsamen Kochen und die Ausflüge, die wir unternommen haben. Ohne euch würde in meinem Leben etwas fehlen.

Ganz besonders bedanke ich mich bei dir, Heike. Du hast mich in die ganzen letzten Wochen und Monaten, sogar Jahren ertragen. Warst immer bereit mir zu helfen, mich zu unterstützen. Du hast mich gefordert, gefördert und motiviert. Ich danke dir und liebe dich!

Lieben Dank auch an meine Familie - meinen Eltern, meinem Bruder und meiner Oma. Ihr habt mir nie einen Weg vorgeschrieben, habt mich finanziell unterstützt und doch ist aus mir etwas Vernünftiges geworden.

Der letzte Dank gilt dem Institut für Mathematik und Informatik. Vielen Dank für die hervorragende Lehre und die familiäre Atmosphäre während meines gesamten Studiums.

Professor Bandt gilt der Dank für die Möglichkeit mit der Arbeitsgruppe zusammenzuarbeiten, für die motivierenden und lobenden Worte sowie die Zuteilung des Themas dieser Diplomarbeit.

Inhaltsverzeichnis

1	Einleitung	8
2	Von der Historie bis zur Gegenwart	10
2.1	Korrelation und Unabhängigkeit	10
2.2	Klassische Verfahren	13
2.2.1	Produkt-Moment-Korrelation von Pearson	13
2.2.2	Rangkorrelation von Spearman	15
2.2.3	Pearsons χ^2 -Test zur Unabhängigkeit	16
2.2.4	Fisher-Yates-Test	18
2.2.5	Kendalls τ	22
2.2.6	Quadrantentest	24
2.2.7	Anwendung der klassischen Verfahren - ein Beispiel	26
2.3	Moderne Verfahren	28
2.3.1	Barnard's CSM-Test	28
2.3.2	Hoeffdings D-Test of Independence	33
2.3.3	Bakirovs I_n	38
2.3.4	Permutation-Entropy-Test von García-Marín	41
2.3.5	Längste aufsteigende Teilfolge L_n	46
2.3.6	Geometrie zufälliger Permutationen	54
3	Vergleich der Unabhängigkeitstests	74
3.1	IndepTestTool - eine neue Toolbox für Matlab	74
3.2	Power Analyse	76
3.2.1	Der unabhängige Fall	76
3.2.2	Der abhängige Fall	78
3.3	Praktische Beispiele	85
4	Zusammenfassung	89
	Bibliographie	90
5	Anhang	94
5.1	Fliegendatensatz	94
5.2	p-Werte des Korrelationsbeispiels	95
5.3	Quantile der Verteilung von Hoeffdings D	96
5.4	Quantile der Verteilung von Bakirovs I_n	97
5.5	Quantile der Teststatistik G des Permutation-Entropy-Tests	98
5.6	Quantile der Länge der längsten aufsteigenden Teilfolge L_n (zweiseitig)	100
5.7	Quantile von S_u und S_f des Geometrietests	101

1 Einleitung

„Vergleich von Unabhängigkeitstests für stetige Größen“, so lautet der Titel dieser Arbeit. Das hört sich beim ersten Lesen vielleicht etwas trocken, langweilig und nach wenig Neuem an. Doch das wird keineswegs der Fall sein, denn es wird uns ein neuer statistischer Unabhängigkeitstest erwarten, der zudem noch gute Eigenschaften besitzt. Ziel soll es dabei sein, bei gegebener Punktwolke möglichst richtig entscheiden zu können, ob zwischen den Merkmalen ein Zusammenhang besteht. Zusammenhänge zu erkennen oder wenigstens zu vermuten ist von großer Wichtigkeit in der Forschungsarbeit auf fast allen Gebieten der Wissenschaft. So können Biologen naturgegebene, biologische oder chemische Abhängigkeiten aufdecken. Ist Enzym A verantwortlich für die Bindung von Substrat 1 und 2? Oder welche Phenole sind für die violette Blütenfärbung des Veilchens verantwortlich? Eine von vielen psychologischen Fragestellungen wäre zum Beispiel: Wirkt sich Mittagsschlaf positiv auf den Stresspegel eines Menschen aus? So gibt es in jedem wissenschaftlichen Gebiet Fragestellungen, zu denen gewisse Merkmale gemessen werden, um diese und jene Schlüsse ziehen zu können. Wir wollen uns dabei auf den Zusammenhang zweier stetiger Variablen konzentrieren. Einige Konzepte, wie das Korrelationsmaß von Pearson, die Rangkorrelation von Spearman oder der χ^2 -Test auf Unabhängigkeit sind uns dabei schon aus dem statistischen Grundstudium bekannt. Darüber hinaus gibt es aber eine Vielzahl von weiteren Tests, die wir schrittweise durcharbeiten und in Matlab programmieren wollen. Neben der Pearson- und Spearmankorrelation [9][10], wie dem χ^2 -Test [12] werden wir als klassische Verfahren den Fisher-Yates-Test, Kendalls τ [15] und den Quadrantentest [23] betrachten. Modernere Ideen brachten BARNARD [17], Hoeffding [18], BAKIROV ET AL. [41], GARCÍA/MARÍN [44] und GARCIA/GONZÁLEZ-LÓPEZ [45]. Interessant wird es mit der Einführung eines neuen Testes, dem GRaP Independence Test. Was sich hinter diesen Buchstaben verbirgt und wie die anderen Tests funktionieren, das sei Gegenstand von Kapitel 2. Im Kapitel 3 wollen wir erörtern, mit welchem Test Zusammenhänge besonders gut erkannt werden können. Dabei werden wir verschiedene Korrelationsmuster betrachten und einige praktische Beispiele durchrechnen. Den Abschluss bildet Kapitel 4, in dem wir die gewonnenen Erkenntnisse dieser Diplomarbeit zusammenfassen.

Hinweise zur beiliegenden CD

Dieser Diplomarbeit ist eine Daten-CD beigelegt (siehe Hardcover letzte Seite). Sie enthält die Diplomarbeit im PDF-Format mit Verlinkungen zu den wissenschaftlichen Publikationen. Weiterhin stelle ich sämtliche programmierten Unabhängigkeitstests bereit. Im Ordner `Matlab` sind diese Programme der Form `*Ind.m`, wobei `*` für den jeweiligen Testnamen steht. Die Routinen zur Berechnung der Verteilungsfunktionen diverser Teststatistiken tragen den Namen `*Df.m`. Die aus diesen Routinen resultierenden Matlabmatrizen sind im Ordner `\Matlab\Matlab.SaveFiles` abgespeichert. Der Abruf der Matrizen erfolgt durch den Import-Befehl von Matlab und kann zur Berechnung der p-Werte verwendet werden. Im Unterordner `Power` befinden sich Matrizen mit p-Werten, die sich aus den verschiedenen Simulationen zur Teststärke ergeben hatten. Im Rahmen der Diplomarbeit ist auch eine neue Toolbox entstanden. Sie trägt den Namen `IndepTestTool` und kann mit eben diesem Namen aufgerufen werden.

2 Von der Historie bis zur Gegenwart

2.1 Korrelation und Unabhängigkeit

Bevor wir uns den Unabhängigkeitstests widmen, beginnen wir mit einer kurzen Reise zu den Wurzeln der so wichtigen Begriffe, wie Korrelation und Unabhängigkeit. Wir starten im 18. Jahrhundert, genau genommen im Jahre 1763, als Richard Price ein Paper seines verstorbenen Freundes Thomas Bayes fand. In *An Essay towards solving a Problem in the Doctrine of Chances* [7] führte Bayes nicht nur den Begriff der bedingten Wahrscheinlichkeit ein und den nach ihm benannten Satz von Bayes, sondern formulierte auch erstmals die Unabhängigkeit zweier Ereignisse:

„Events are independent when the happening of any one of them does neither increase nor abate the probability of the rest.“

Die wahrscheinlich erste mathematische Definition der Unabhängigkeit stammt allerdings von Pierre-Simon Laplace. Er wies darauf hin, dass, wenn $\{E_i\}$ eine Folge unabhängiger Ereignisse ist mit $p_i = P(E_i)$, so ist $P(E_1, \dots, E_n) = \prod_{i=1}^n p_i$. Weiterhin bemerkte er für den Fall zweier Ereignisse, dass $P(E_1 E_2) = P(E_2 | E_1) P(E_1)$ unter der Annahme, dass der Eintritt des ersten Ereignisses E_1 die Wahrscheinlichkeit des Eintretens des zweiten Ereignisses E_2 beeinflusst. Nach heutiger mathematischer Notation (aus [5]) würden wir die Unabhängigkeit zweier Ereignisse so formulieren : Seien $A, B \subset \Omega$ zwei Ereignisse. A und B heißen stochastisch unabhängig, wenn gilt:

$$P(A \cap B) = P(A) \cdot P(B) \quad \text{bzw.}$$

$$P(A|B) = P(A) \text{ mit } P(B) > 0 \quad \text{bzw.}$$

$$P(B|A) = P(B) \text{ mit } P(A) > 0$$

Die Unabhängigkeit zweier Zufallsvariablen definiert sich analog zur stochastischen Unabhängigkeit von Ereignissen. Die Zufallsvariablen X und Y heißen unabhängig, wenn für

alle x und y gilt:

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

Ansonsten heißen X und Y abhängig. Eine Zufallsvariable X sei dabei eine auf die Ergebnismenge Ω definierte Abbildung, die jedem $\omega \in \Omega$ eine reelle Zahl $X(\omega)=x$ zuordnet. Ein jüngeres Konzept mit einer schwächeren Eigenschaft für die Unabhängigkeit ist die Korrelation. Dessen Wurzeln gehen zurück auf Carl Friedrich Gauß, der 1823 den Normalenvektor von n korrelierten Zufallsvariablen definierte. Doch Gauß interessierte sich wohl nicht für die einzelnen Beziehungen seiner Terme, so Pearson in *Notes on the History of Correlation* [11]. Auch Auguste Bravais berichtete von den Parametern der bivariaten Normalverteilung. Er führte zwar die Produkt-Summe ein, aber kein Symbol für den Korrelationskoeffizienten. Ähnlich wie Gauß erkannte er nicht die Wichtigkeit der Korrelation als Maß des Zusammenhangs von Variablen. Es folgte Sir Francis Galton, ein Halbcousin Darwins, britischer Naturforscher und Schriftsteller des 19. Jahrhunderts, der in seinem Werk *Regression Towards Mediocrity in Hereditary Stature* [8] die Regression als praktische Methode entwickelte, um einen Zusammenhang zwischen der Körpergröße von Eltern und Kind zu beweisen. Galton veröffentlichte dazu auch den ersten Graph zur Korrelation:

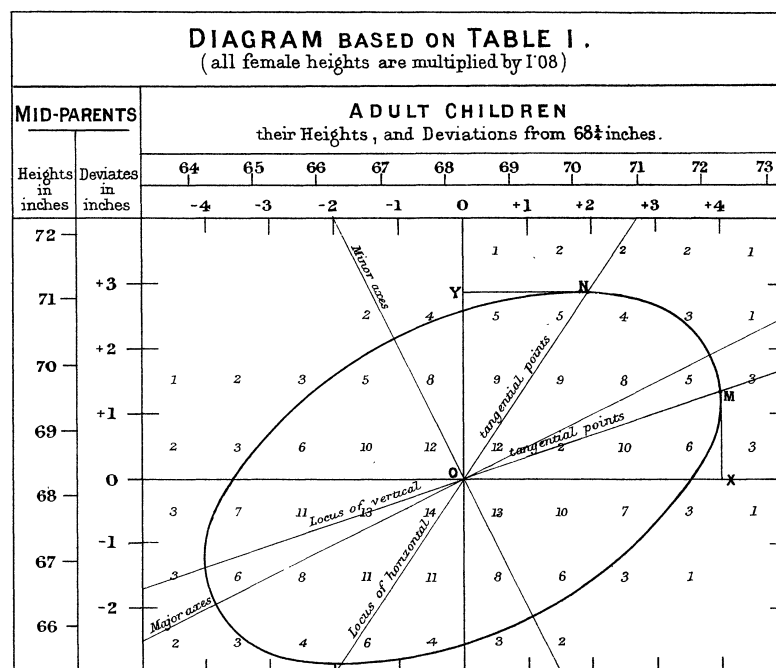


Abbildung 2.1: Das erstes bivariate Streudiagramm von Galton 1885 (aus [8]).

2 Von der Historie bis zur Gegenwart

Jenes Werk war es auch, welches die Vorlage für die Definition des Korrelationskoeffizienten lieferte, den sein Freund Karl Pearson schließlich 1895 mathematisch manifestierte. Und so wird jetzt endlich Sir Francis Galton als Vater der Korrelation bezeichnet.

Meilensteine in der Geschichte von Korrelation und Regression (Original von Rodgers/Nicewander [32])		
Jahr	Autor	Ereignis
1823	Carl Friedrich Gauß, deutscher Mathematiker	Entwickelte den Normalenvektor von n korrelierten Zufallsvariablen.
1843	John Stuart Mill, britischer Philosoph und Ökonom	Stellte Regeln zur induktiven Logik auf, einschließlich der Methode der gleichzeitigen Änderungen.
1846	Auguste Bravais, französischer Marineoffizier und Astronom	Arbeitete an der bivariaten Normalverteilung und erwähnte „une correlation“.
1868	Charles Darwin, Galtons Cousin, britischer Naturforscher	„All parts of the organisation are [...] connected or correlated.“
1877	Sir Francis Galton, Brite, erster Biometriker	Erste Erörterung von <i>Reversion</i> , dem Vorgänger der Regression.
1885	Sir Francis Galton	Erste Erwähnung des Regressionsbegriffes Veröffentlichung des bivariaten Streudiagramms mit Linien gleicher Wahrscheinlichkeitsdichte, der erste Graph zur Korrelation. Fertigstellung der Theorie von der bivariaten Korrelation.
1888	Sir Francis Galton	Konzeptionelle Definition von r , Angabe der oberen Grenze.
1895	Karl Pearson, britischer Statistiker	Definition des (Galton-) Pearson Produkt-Moment-Korrelationskoeffizienten.
1920	Karl Pearson	Veröffentlichung von <i>Notes on the History of Correlation</i> [11]
2010		125 Jahre Regression und Korrelation

Bei der Definition des Korrelationskoeffizienten von Pearson wollen wir nun ansetzen. Auf den folgenden Seiten werden wir weitere Methoden und Tests kennen lernen, um entscheiden zu können, ob eine Abhängigkeit zweier Variablen vorliegt.

2.2 Klassische Verfahren

2.2.1 Produkt-Moment-Korrelation von Pearson

Wie wir bereits erfahren haben, führte SIR FRANCIS GALTON 1885 den Begriff der Regression ein [8]. Nachdem er 1888 erkannte, dass r als Maß zur Korrelation dient und obwohl der Naturforscher bemerkte, dass r nicht größer als 1 sein kann, entwickelte sein Freund KARL PEARSON [9] erst 7 Jahre später das mathematische Fundament zum bekanntesten und am häufigsten genutzten Koeffizienten zur Messung des linearen Zusammenhangs zweier Merkmale.

In vielen statistischen Büchern wird der Korrelationskoeffizient r als Produkt-Moment-Korrelation wie folgt definiert:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.1)$$

Die x_1, \dots, x_n und y_1, \dots, y_n seien dabei die empirischen Ausprägungen der Merkmale X und Y . Mit Hilfe der Cauchy-Schwarzschen Ungleichung kann gezeigt werden, dass $r \in [-1, 1]$. Darüber hinaus finden sich diverse äquivalente Definitionen und Interpretationen des Koeffizienten in der lesenswerten Veröffentlichung *Thirteen Ways to Look at the Correlation Coefficient* von RODGERS und NICEWANDER [32], z.B. der Korrelationskoeffizient als standardisierte Kovarianz:

$$r = \frac{s_{XY}}{s_X s_Y}$$

Um in praktischen Bereichen den Zugang zur Auswertung des Koeffizienten zu erleichtern, wird, wie hier entnommen aus *Statistik - Der Weg zur Datenanalyse* [5], eine Einteilung in in folgende Intensitätsgrade vorgenommen:

„schwache Korrelation“	$ r < 0.5$
„mittlere Korrelation“	$0.5 \leq r < 0.8$
„starke Korrelation“	$0.8 \leq r $

Stochastische Unabhängigkeit beider Merkmale ist nur bei $r=0$ möglich. Die Rückrichtung existiert allerdings nicht, denn es gibt einige Fälle, bei denen $r=0$ ist und trotzdem einen starken Zusammenhang zwischen den Merkmalen besteht.

Setzt man also Unabhängigkeit voraus, so legt man die Nullhypothese $H_0 : r=0$ zugrunde.

2 Von der Historie bis zur Gegenwart

Wenn weiterhin X und Y normalverteilte Stichprobenvariablen sind, so kann man die Teststatistik T anlegen:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (2.2)$$

Unter H_0 sei die Teststatistik T t -verteilt mit $n-2$ Freiheitsgraden. $t_{1-p/2}(n-2) = T$ ergibt das zweiseitige Signifikanz-Niveau p für den Korrelations-Test.

Matlab [Pearson-Korrelation]

Die Berechnung von r ist bereits in Matlab implementiert. Nach Eingabe der Stichprobenvektoren x und y kann daher die unten stehende Funktion verwendet werden. Die Spaltenvektoren müssen dabei die gleiche Dimension besitzen. Als Ausgabe wird die Korrelationsmatrix geliefert.

```
1 %% Pearsonkorrelation
2 corrcoef(x,y)
```

2.2.2 Rangkorrelation von Spearman

Der nach CHARLES SPEARMAN benannte Koeffizient von 1904 [10] stellt den ersten rangbasierten Koeffizienten dar und wird manchmal mit ρ bezeichnet (hier: r_S). Zur Berechnung werden von den ursprünglichen Werten x, y Ränge gebildet:

$$\begin{aligned} \text{rg}(x_i) &= i && \text{für geordnete Werte } x_1 \leq x_2 \leq \dots \leq x_n \\ \text{rg}(y_i) &= i && \text{für geordnete Werte } y_1 \leq y_2 \leq \dots \leq y_n \end{aligned} \quad (2.3)$$

Gibt es identische Werte x_j, \dots, x_{j+k} , genannt *Bindungen* bzw. *ties*, so ist diesen Werten der Durchschnittsrang zuzuweisen:

$$\text{rg}(x_j) = \dots = \text{rg}(x_{j+k}) := \frac{\text{rg}(x_j) + \dots + \text{rg}(x_{j+k})}{k+1}$$

Die Berechnung der Spearmankorrelation wird analog zur Pearsonkorrelation (2.1) durchgeführt, jedoch mit den Paaren $(\text{rg}(x_i), \text{rg}(y_i))_{i=1, \dots, n}$ als Ausprägungen:

$$r_S = \frac{\sum_{i=1}^n (\text{rg}(x_i) - \overline{\text{rg}_x})(\text{rg}(y_i) - \overline{\text{rg}_y})}{\sqrt{\sum_{i=1}^n (\text{rg}(x_i) - \overline{\text{rg}_x})^2 \sum_{i=1}^n (\text{rg}(y_i) - \overline{\text{rg}_y})^2}} \quad (2.4)$$

mit $-1 \leq r_S \leq 1$

$\overline{\text{rg}_x}$ und $\overline{\text{rg}_y}$ bezeichnen dabei die Durchschnittsränge. Durch diese Methodik gelingt es, nicht nur den linearen Zusammenhang zu erfassen, sondern auch jegliche monotone Zusammenhänge. Falls keine *ties* existieren, kann die vereinfachte Formel genutzt werden:

$$r_S = 1 - \frac{6 \sum_{i=1}^n (\text{rg}(x_i) - \text{rg}(y_i))^2}{(n^2 - 1)n}$$

Matlab [Spearman-Korrelation]

Die Berechnung von r_S erfolgt durch folgende Funktion nach Eingabe der Stichprobenvektoren x und y :

```
1 %% Spearmankorrelation
2 [r, p] = corr(x, y, 'type', 'Spearman');
```

2.2.3 Pearsons χ^2 -Test zur Unabhängigkeit

PEARSON führte 1922 einen weiteren statistischen Test [12] ein, welcher gegenwärtig Bestandteil des Grundstudiums vieler Naturwissenschaftler ist. Beim χ^2 -Test seien die Merkmale X und Y diskret. Stetige Merkmale sind daher vorher zu diskretisieren. Die gemeinsame Verteilung kann dann in eine $k \times m$ -Kontingenztabelle geschrieben werden.

Tabelle 2.1: $k \times m$ -Kontingenztabelle

X	Merkmal Y						Summe
	1	2	...	j	...	m	
1	n_{11}	n_{12}	...	n_{1j}	...	n_{1m}	$n_{1\bullet}$
2	n_{21}	n_{22}	...	n_{2j}	...	n_{2m}	$n_{2\bullet}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{im}	$n_{i\bullet}$
\vdots	\vdots	\vdots		\vdots		\vdots	\vdots
k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{km}	$n_{k\bullet}$
Summe	$n_{\bullet 1}$	$n_{\bullet 2}$...	$n_{\bullet j}$...	$n_{\bullet m}$	n

Die Häufigkeiten, die wir bei Unabhängigkeit beider Merkmale zu erwarten haben, berechnen wir anhand der Randhäufigkeiten: $n'_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$. Es ist anzunehmen, dass die Differenz zwischen den beobachteten Häufigkeiten $(n_{ij})_{i=1, \dots, k}$ mit erwarteten Häufigkeiten $(n'_{ij})_{i=1, \dots, k, j=1, \dots, m}$ Null ist, falls die Merkmale wirklich unabhängig sind. Dies führt auf folgende Prüfgröße:

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}} \tag{2.5}$$

Da die Differenzen $n_{ij} - n'_{ij}$ quadriert und normiert werden, sind die Summanden in Gleichung 2.5 als standardnormalverteilte Zufallsgrößen anzusehen. Geht man von stochastischer Unabhängigkeit dieser Größen aus, so ist die Aufsummierung solcher Zufallsgrößen bekannterweise χ^2 -verteilt. Weil zur Berechnung der Prüfgröße die erwarteten Häufigkeiten geschätzt werden müssen, ergeben sich $(k-1)(m-1) = df$ Freiheitsgrade. Der χ^2 -Unabhängigkeitstest ist demnach ein χ^2 -Anpassungstest, bei dem überprüft wird, ob die Beobachtung die Wahrscheinlichkeitsverteilung F_0 besitzt, wobei F_0 die Verteilung bei Unabhängigkeit bezeichnet. Der Signifikanzwert ergibt sich aus:

$$p = P(\chi^2_{((k-1)(m-1))} > \chi^2) \tag{2.6}$$

PEARSON gibt in [12] S.189 eine vereinfachte Formel des χ^2 -Wertes für 2×2 -Tafeln an¹. Dabei bezeichnen a, b, c und d die Einträge n_{11}, n_{12}, n_{21} und n_{22} :

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(a + c)(b + c)(c + d)} \quad (2.7)$$

Matlab [χ^2 -Test auf Unabhängigkeit]

Die Berechnung der Prüfgröße und des p-Wertes erfolgt durch folgende Funktion nach Eingabe der Stichprobenvektoren x und y :

```

1 function p=Chi2Ind(x,y)
2 %Syntax: p=Chi2Ind(x,y)
3 % Calculates p-value by testing independence with the Chi-square test of
4 % independence by Karl Pearson on given data-vectors x and y.
5 %
6 % Marcus Vollmer
7 % 16.05.2010
8
9 S(4)=zeros;
10 n=size(x,1);
11
12 %Anzahlen in den Quadranten
13 for i=1:n
14     if x(i)<median(x)
15         if y(i)>=median(y)
16             S(1)=S(1)+1;
17         else
18             S(3)=S(3)+1;
19         end
20     else if y(i)>=median(y)
21         S(2)=S(2)+1;
22     else
23         S(4)=S(4)+1;
24     end
25 end
26 end
27
28 chi=(sum(S)*(S(1)*S(4)-S(3)*S(2))^2)/...
29     ((S(1)+S(2))*(S(1)+S(3))*(S(2)+S(4))*(S(3)+S(4)));
30
31 p=1-chi2cdf(chi,1);
32 end

```

¹Pearson muss sich dabei in seinem Paper [12] auf den Seiten 189 und 190 verdrückt haben. Er vertauschte im Nenner von Gleichung (2.7) b und d . Die hier abgedruckte Formel ist richtig angegeben und ist durch Umformungen von (2.5) herzuleiten.

2.2.4 Fisher-Yates-Test

Der Fisher-Yates-Test aus dem Jahre 1925 wird vor allem dann angewendet, falls zwei dichotome Merkmale vorliegen. Dies ist bei vielen biologischen und medizinischen Untersuchungen der Fall, bei denen Merkmale wie das Geschlecht (männlich/weiblich) oder typische Ja-Nein-Fragen (Raucher/Nichtraucher) analysiert werden. Die Kombinationen der Merkmalsausprägungen werden in der Regel in eine 2×2 -Kontingenztafel eingetragen. Deren Häufigkeiten seien mit a_0, b_0, c_0, d_0 bezeichnet und die Tabelle ist so anzuordnen, dass $a_0 \leq b_0, c_0, d_0$.

	A	\bar{A}	Summe
B	a_0	b_0	$a_0 + b_0$
\bar{B}	c_0	d_0	$c_0 + d_0$
Summe	$a_0 + c_0$	$b_0 + d_0$	n

Geht man von Unabhängigkeit beider Merkmale aus und interpretiert die Anzahlen a_0 und b_0 als gezogene schwarze und weiße Kugeln einer Urne mit insgesamt $a_0 + c_0$ schwarzen und $b_0 + d_0$ weißen Kugeln, so lässt sich anhand dieses Urnenmodells die Wahrscheinlichkeit p_0 der Beobachtung berechnen.

$$p_0 = \frac{\binom{a_0 + c_0}{a_0} \binom{b_0 + d_0}{b_0}}{\binom{n}{a_0 + b_0}} = \frac{(a_0 + c_0)! (b_0 + d_0)! (a_0 + b_0)! (c_0 + d_0)!}{a_0! b_0! c_0! d_0! n!} \quad (2.8)$$

Beobachtungen, die stärker gegen die Nullhypothese, der Unabhängigkeit der Merkmale, sprechen, werden als extremere Stichproben bezeichnet. Diese erhält man, wenn a_0 einen noch geringeren Wert annimmt, den man mit $a_i = a_0 - i$ für $i = 1, \dots, a_0$ bezeichnet. Bei festen Randhäufigkeiten ergeben sich die restlichen Werte durch Vervollständigung der Tabelle: $b_i = b_0 + i$, $c_i = c_0 + i$ und $d_i = d_0 - i$. Die Wahrscheinlichkeiten der Stichproben $(a_i, b_i, c_i, d_i)_{i=1, \dots, a_0}$ seien mit $(p_i)_{i=1, \dots, a_0}$ bezeichnet und lassen sich rechenstechnisch sinnvoll nach FELDMAN, KLINGER [24] rekursiv berechnen:

$$p_{i+1} = \frac{a_i d_i}{b_{i+1} c_{i+1}} p_i \quad (2.9)$$

Die Summe $p=p_0+\dots+p_{a_0}$ liefert den einseitigen Signifikanzwert des Fisher-Yates-Tests. Zur zweiseitigen Prüfung gelten zudem jene Beobachtungen als extremer, für die gilt.

$$\left| \frac{a}{a+b} - \frac{c}{c+d} \right| \geq \left| \frac{a_0}{a_0+b_0} - \frac{c_0}{c_0+d_0} \right| \quad (2.10)$$

Aufgrund des Aufwandes zur Berechnung der Fakultäten in (2.8) wird der Test i.d.R. nur für einen kleinen Stichprobenumfang verwendet. Dafür erfolgt die Berechnung des p -Wertes exakt. Für größere Stichprobenumfänge stößt der Computer zur Berechnung von (2.8) an die Grenzen der Zahlendarstellung und deren Genauigkeit. Hier sollte meiner Ansicht nach, eine effizientere Berechnung verfolgt werden. Durch Bildung des Logarithmus naturalis vereinfacht sich Formel (2.8) zu

$$\begin{aligned} \ln(p_0) = & \ln((a_0+c_0)!) + \ln((b_0+d_0)!) + \ln((a_0+b_0)!) + \ln((c_0+d_0)!) \\ & - \ln(a_0!) - \ln(b_0!) - \ln(c_0!) - \ln(d_0!) - \ln(n!). \end{aligned} \quad (2.11)$$

In analoger Weise vereinfacht sich Formel (2.9) zu

$$\ln(p_{i+1}) = \ln(a_i) + \ln(d_i) - \ln(b_{i+1}) - \ln(c_{i+1}) + \ln(p_i). \quad (2.12)$$

Das Signifikanz-Level ergibt sich dann durch $p = \exp(\ln(p_0)) + \dots + \exp(\ln(p_{a_0}))$. Die Berechnung der Logarithmen von Fakultäten der Art $\ln(n!)$, wie in (2.11), kann durch die Stirling-Formel approximiert werden:

$$\ln(n!) = \frac{1}{12n} + \frac{\ln(2\pi)}{2} + (n+1/2) \ln(n) - n. \quad (2.13)$$

Auf diese Weise gelingt eine schnelle und zuverlässige Berechnung auch für größere Stichproben. Doch nicht nur bei dichotomen Merkmalen kann dieser Test nützlich sein. Stetige Merkmale können z.B. durch Klassenbildung mittels Median ($A=(X \leq X_{\text{med}})$, $\bar{A}=(X > X_{\text{med}})$) zu dichotomen Merkmalen transformiert werden.

Darüber hinaus gibt es eine Verallgemeinerung auf $k \times m$ -Tafeln von FREEMAN und HALTON [19], welche auch von BORTZ, LIENERT & BOEHNKE [6] beschrieben wurde, sowie einen weiteren exakten Test aus den 40er Jahren von G. A. BARNARD [16],[17], den wir im Kapitel 2.3.1 betrachten werden.

Matlab [Fisher-Yates-Test]

Die Berechnung des p -Wertes erfolgt nach Eingabe der Stichprobenvektoren x und y . Das Koordinatensystem wird dabei durch die Mediane in vier Quadranten geteilt. Die

2 Von der Historie bis zur Gegenwart

Häufigkeiten a, b, c, d ergeben sich dabei aus der Anzahl der Punkte in den Quadranten. Für $n \geq 50$ wird approximativ gerechnet. Dabei werden Logarithmen genutzt und die Stirling-Formel verwendet.

```
1 function p=FisherYatesInd(x,y)
2 %Syntax: p=FisherYatesInd(x,y)
3 % Calculates p-value by testing Independence with Fisher-Yates (1925)
4 %
5 % Inputs: x - nx1 vector of data
6 %         y - nx1 vector of data
7 % Output:
8 %         p - significance level
9 %
10 % Marcus Vollmer
11 % 07.05.2010 / 02.06.2010
12
13 S1=0; S2=0; S3=0; S4=0;
14 n=size(x,1);
15
16 %quadrantscores
17 for i=1:n
18     if x(i)>=median(x)
19         if y(i)>=median(y)
20             S1=S1+1;
21         else
22             S4=S4+1;
23         end
24     else if y(i)>=median(y)
25         S2=S2+1;
26     else
27         S3=S3+1;
28     end
29 end
30 end
31
32 %definition of a,b,c,d
33 if S2<=S1 && S2<=S3 && S2<=S4
34     a=S2; b=S1; c=S3; d=S4;
35 else if S1<=S3 && S1<=S4
36     a=S1; b=S2; c=S4; d=S3;
37 else if S3<=S4
38     a=S3; b=S4; c=S2; d=S1;
39 else
40     a=S4; b=S3; c=S1; d=S2;
41 end
42 end
43 end
44
45 %calculating significance level
46 P=zeros(a+1,1);
47 if n<50 %using nchoosek function
```

```

48     P(1)=nchoosek(a+c,a)*nchoosek(b+d,b)/nchoosek(n,a+b);
49     k=a+1;
50     for i=2:k
51         b=b+1; c=c+1;
52         P(i)=(a*d*P(i-1))/(b*c);
53         a=a-1; d=d-1;
54     end
55     p=sum(P);
56     else %using approximation of the P(i)'s for large samples
57     P(1)=lnapprox(a+c)+lnapprox(b+d)+lnapprox(a+b)+lnapprox(c+d)-lnapprox(a
58         )-lnapprox(b)-lnapprox(c)-lnapprox(d)-lnapprox(n);
59     k=a+1;
60     for i=2:k
61         b=b+1; c=c+1;
62         P(i)=log(a)+log(d)+P(i-1)-log(b)-log(c);
63         a=a-1; d=d-1;
64     end
65     p=sum(exp(P));
66 end
67
68 function lnfac=lnapprox(l)
69 % Calculation of ln(l!) in asymptotic way using Stirlings approximation
70 if l==0
71     lnfac=0;
72 else
73     lnfac=1/(12*l)+log(2*pi)/2+(1+1/2)*log(l)-1;
74 end
75 end

```

2.2.5 Kendalls τ

SIR MAURICE GEORGE KENDALL präsentierte der Öffentlichkeit 1938 einen weiteren rangbasierten Koeffizienten [15]. Dieser ist nach dem Briten benannt und wird mit dem griechischen Buchstaben τ bezeichnet. Bei genauerer Betrachtung ist theoretisch keine Rangbildung nötig, denn zur Berechnung dieser Kennziffer ist lediglich die Lage der Wertepaare (X, Y) interessant. Bei monotonen Zusammenhängen, wie in Abbildung 2.2, liegen die Wertepaare, ausgehend von dem Punkt (x_i, y_i) , jeweils nur in den Quadranten 1 und 3. Bei weniger starken Zusammenhängen in allen vier Quadranten. Als Maß für den Zusammenhang subtrahiert man die Anzahl der Punkte in den Quadranten 2 und 4 von der Anzahl an Punkten in den Quadranten 1 und 3. Führt man dies für alle Paare $(x_i, y_i)_{i=1, \dots, n}$ durch und addiert dessen Werte, so erhält man den Stichprobenscore S . Der maximal mögliche Score ist $n(n-1)$ und tritt bei einem monotonen Zusammenhang auf. Kendalls τ ergibt sich somit als Quotient des Stichprobenscores mit dem maximalen Score und befindet sich trivialerweise zwischen -1 und 1 :

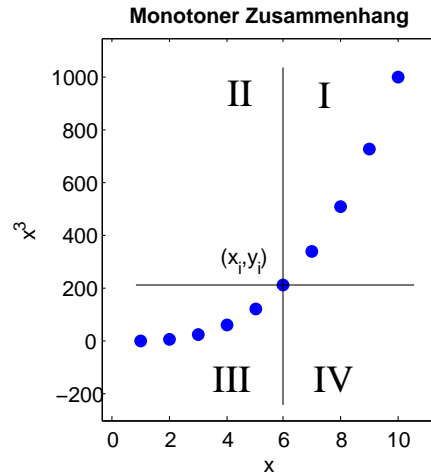


Abbildung 2.2: Lage bei monotonen Zusammenhängen.

$$\tau = \frac{\text{Stichprobenscore}}{\text{Maximal möglicher Score}} = \frac{\sum_{i=1}^n \sum_{j \neq i} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{n(n-1)} \quad (2.14)$$

Um zu überprüfen, ob sich das Ergebnis signifikant von der Unabhängigkeit unterscheidet, kann für einen kleinen Stichprobenumfang die Verteilung der Scores genau berechnet werden [vgl. Abbildung 2.3]. Wie gewöhnlich kann dann ein Normal- und Extrembereich festgelegt werden. Für große n können wir den p -Wert anhand der Normalverteilung berechnen, denn bei zufälliger Anordnung der Punkte ohne „ties“ ist der Stichprobenscore S normalverteilt um 0 mit einer Varianz von $\sigma^2(S) = \frac{1}{18} n(n-1)(2n+5)$.

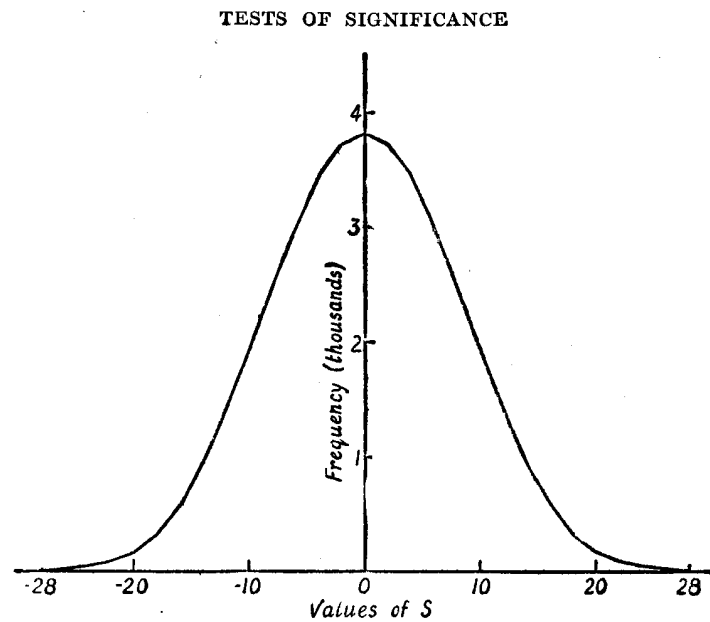


Abbildung 2.3: Verteilung für $n=8$. Bei $\alpha=5\%$ ergibt sich ein Normalbereich von $[-16; 16]$.
Abbildung entnommen aus Kapitel 4, *Rank Correlation Methods* [2].

Aufgrund dessen, dass S nur diskrete Werte annehmen kann, erfolgt die Berechnung des p -Wertes nach einer Stetigkeitskorrektur (engl. correction for continuity) wie folgt:

$$p = \begin{cases} 2(1 - \Phi(\frac{S-1}{\sigma(S)})), & \text{falls } S > 0, \\ 1, & \text{falls } S = 0, \\ 2\Phi(\frac{S+1}{\sigma(S)}), & \text{falls } S < 0. \end{cases} \quad (2.15)$$

Matlab [Kendalls τ]

Die Berechnung von τ erfolgt durch die gegebene Funktion nach Eingabe der Stichprobenvektoren x und y :

```

1 %% Kendalls Tau
2 [tau,p]=corr(x,y,'type','Kendall');
```

2.2.6 Quadrantentest

REGINA C. ELANDT führte Mitte der fünfziger Jahre diesen einfachen, aber weniger bekannten Test ein, welcher im Englischen mit den drei Namen „quadrant measure of association“, „medial correlation“ und „measure of tendency“ benannt ist [vgl. R. C. ELANDT [23]]. Getestet wird wieder die Nullhypothese H_0 , dass die Zufallsvariablen X und Y unabhängig sind. Dazu werden die vier Quadranten betrachtet, welche durch die Mediane von X und Y getrennt werden. Man geht davon aus, dass bei Unabhängigkeit die Anzahl der Datenpunkte in den Quadranten gleich groß sind. Weicht die Anzahl der Datenpunkte in den betrachteten Quadranten signifikant ab, ist eine Abhängigkeit nachgewiesen. Die Teststatistik sei mit S bezeichnet:

$$S = S_1 + S_2 + S_3 + S_4 \quad (2.16)$$

wobei:

$$\begin{aligned} S_1 &= \#(x_j, y_j), \text{ falls } x_j > x_{\text{med}}, y_j > y_{\text{med}} \\ S_2 &= 1/2, \text{ falls } \exists j : x_j = x_{\text{med}}, y_j > y_{\text{med}}, \text{ sonst } S_2 = 0 \\ S_3 &= 1/2, \text{ falls } \exists j : x_j > x_{\text{med}}, y_j = y_{\text{med}}, \text{ sonst } S_3 = 0 \\ S_4 &= 1/4, \text{ falls } \exists j : x_j = x_{\text{med}}, y_j = y_{\text{med}}, \text{ sonst } S_4 = 0 \end{aligned}$$

$\#(x_j, y_j)$ sei dabei die Anzahl der Paare (x_j, y_j) , für die die obige Bedingung gilt. x_{med} und y_{med} seien die beiden Mediane. Ist die Anzahl der n Wertepaare gerade, so ist $S = S_1$. Unter H_0 ist S asymptotisch normalverteilt um den Erwartungswert $\mathbb{E}(S) = \frac{1}{4}n$ bei einer Varianz von

$$\text{Var}(S) = \begin{cases} \frac{1}{16} \frac{n^2}{n-1}, & \text{für } n \text{ gerade} \\ \frac{1}{16}(n-1), & \text{für } n \text{ ungerade} \end{cases} \quad (2.17)$$

Matlab [Quadrantentest]

Die Berechnung von r_S erfolgt durch folgende Funktion nach Eingabe der Stichprobenvektoren x und y :

```

1 function p=QuadrantInd(x,y)
2 %Syntax: p=QuadrantInd(x,y)
3 % Calculates p-value by testing independence by the Quadrant Test by Elandt
4 % on given data-vectors x and y.
```



```

5 %
6 % Marcus Vollmer
7 % 26.04.2010
8
9 n=size(x,1);
10 S1=0; S2=0; S3=0; S4=0;
11 x_med=median(x);
12 y_med=median(y);
13
14 for i=1:n
15     if x(i)>=x_med && y(i)>=y_med
16         S1=S1+1;
17         if x(i)==x_med
18             if y(i)==y_med
19                 S4=1/4; S1=S1-1;
20             else S2=1/2; S1=S1-1;
21             end
22         end
23         if y(i)==y_med && x(i)>x_med
24             S1=1/2; S1=S1-1;
25         end
26     end
27 end
28
29 S=S1+S2+S3+S4;
30
31 %cdf of normal distribution
32 if rem(n,2)==0
33     p=normcdf(S,n/4,(n*n)/(16*(n-1)));
34 else
35     p=normcdf(S,n/4,(n-1)/16);
36 end
37
38 %two-sided p-value
39 if p>0.5
40     p=2*(1-p);
41 else p=2*p;
42 end
43 end

```

2.2.7 Anwendung der klassischen Verfahren - ein Beispiel

Zur Anwendung der klassischen Verfahren verwenden wir den frei erhältlichen Fliegendatensatz² aus dem Datensatz-Archiv des Instituts für Statistik der Ludwig-Maximilians-Universität München und des Sonderforschungsbereichs 386. Gemessen wurden die Körpermaße und die Geweihlängen einer Fliegenart. Die Daten befinden sich auch im Anhang auf Seite 94. Uns soll an dieser Stelle der Zusammenhang zwischen der Körperlänge x und der Geweihlänge y einer Fliege interessieren.

Matlab [Auswertung Fliegendatensatz]

Die Auswertung des Datensatzes verläuft in 3 Schritten. Zuerst werden die erforderlichen Daten geladen und in den Variablen x und y gespeichert. Dies geschieht unter Verwendung von `textread` in der Zeile 5 des Programmes. Die Berechnungen der p-Werte und der Korrelationskoeffizienten werden nach den zuvor definierten Matlab-Funktionen vollzogen (Zeilen 8-13). Als letzter Schritt erfolgt die Ausgabe der Datenpunkte in einem Scatterplot und daneben die der zugehörigen Koeffizienten (Zeilen 15 ff.).

```

1 %% Beispiel Fliegendaten
2 % Laden des Datensatzes
3 % x - Körperlänge einer Fliege
4 % y - Geweihlänge einer Fliege
5 [x, muell, muell, y, muell] = textread('Data\fliegen.asc', '%f %s %s %f %s',
    , 'headerlines', 1, 'whitespace', '', 'delimiter', '\t');
6
7 %Berechnung der Koeffizienten und P-Werte
8 [r, p_pearson]=corrcoef(x,y); %Pearsoncorrelation
9 [r_s, p_spearman]=corr(x,y, 'type', 'Spearman'); %Spearmancorrelation
10 [p_chi2]=Chi2Ind(x,y); %Chi^2 Independence Test
11 [p_fisher]=FisherYatesInd(x,y); %Fisher-Yates-Test
12 [tau, p_kendall]=corr(x,y, 'type', 'Kendall'); %Kendalls Tau
13 [p_quadranten]=QuadrantInd(x,y); %Quadrant Test
14
15 %Plot
16 figure(1)
17 scatter(x,y, 'filled')
18 set(gcf, 'PaperSize', [20 11], 'position', [100 100 800 400], 'color', [1 1 1]);
19 set(gca, 'xlim', [1.15*min(x)-0.15*max(x) 1.15*max(x)-0.15*min(x)], 'ylim',
    , [1.15*min(y)-0.15*max(y) 1.15*max(y)-0.15*min(y)], 'box', 'on')
20 set(gca, 'OuterPosition', [0 0 0.7 1]);
21 title('\fontsize{12}Fliegendatensatz', 'FontWeight', 'bold');
22 xlabel('\fontsize{12}Körperlänge');
23 ylabel('\fontsize{12}Geweihlänge');
24

```

²<http://www.statistik.lmu.de/service/datenarchiv/fliegen/fliegen.html>

2.2.7 Anwendung der klassischen Verfahren - ein Beispiel

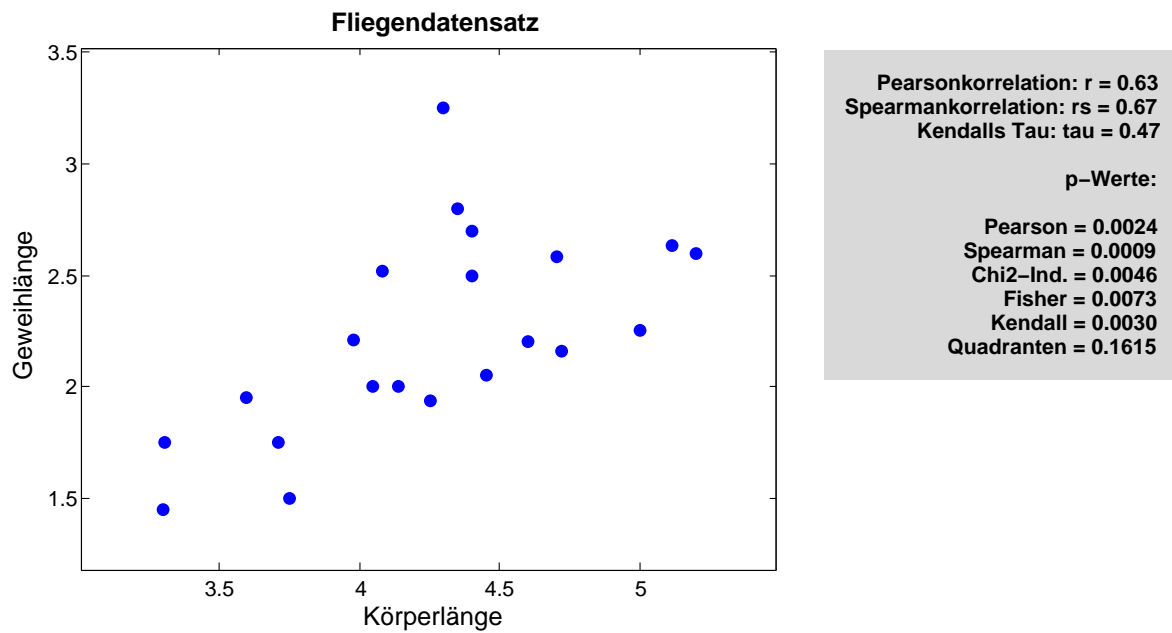


Abbildung 2.4: Streudiagramm, Korrelations- und Signifikanzwerte des Fliegendatensatzes.

```

25 text(max(x)+6*(max(x)-min(x))/7,max(y)+(max(y)-min(y))/11,sprintf('
    Pearsonkorrelation: r = %.2f\nSpearmankorrelation: rs = %.2f\nKendalls
    Tau: tau = %.2f\n\np-Werte:\n\nPearson = %.4f\nSpearman = %.4f\nChi2-
    Ind. = %.4f\nFisher = %.4f\nKendall = %.4f\nQuadranten = %.4f',r(2), r_s
    , tau, p_pearson(2), p_spearman, p_chi2, p_fisher, p_kendall,
    p_quadranten),'FontWeight','Bold','HorizontalAlignment','right','
    VerticalAlignment','top','BackgroundColor',[0.85 0.85 0.85],'Margin',10)
    ;

```

Das Streudiagramm zeigt uns einen mittleren linearen Zusammenhang. Pearson und Spearman bestätigen diesen Zusammenhang durch die Koeffizienten $r=0.63$ und $r_s=0.67$. Die klassischen Tests auf Unabhängigkeit liefern hohe Signifikanzwerte von 0.0009 bis 0.0073. Das heißt, es ist höchst unwahrscheinlich, dass die beiden Merkmale Körperlänge und Geweihlänge unabhängig sind. Eine stochastische Abhängigkeit gilt damit als gesichert. Lediglich der Quadrantentest scheitert und weist keinen Zusammenhang auf.

2.3 Moderne Verfahren

In diesem Abschnitt sollen neuere Verfahren und Ideen vorgestellt werden. Wir werden feststellen, dass viele Ideen leicht verständlich und dennoch komplexer sind.

2.3.1 Barnard's CSM-Test

Gleich seine erste Publikation *A New Test for 2×2 Tables* erschien 1945 in der Fachzeitschrift *Nature* [16]. Der Statistiker George Alfred Barnard behauptete, er habe einen neuen exakten statistischen Test zur Unabhängigkeit entwickelt, der eine größere Power besitzt, als der exakte Test von Fisher³ (siehe 2.2.4, Seite 18). Auf dem halbseitigen Artikel beschreibt Barnard nur in Auszügen seine Ideen und Folgerungen und verschiebt die Details auf spätere Publikationen. Barnard weiß zu berichten, dass Fisher über die Degradierung seines Testes nicht erfreut war und in einem Brief an ihn gegen den neuen Test wettete⁴. Nach Briefwechseln und persönlichen Treffen wurden die beiden aber schließlich Freunde. Barnard lieferte die genauere mathematische Beschreibung dann zwei Jahre später, zu finden in der *Biometrika: Significance Tests for 2×2 Tables* [17]. Auf den nächsten Seiten soll der Test im Detail beschrieben werden.

Das Modell

Gegeben sei folgende Kontingenztafel:

	<i>A</i>	<i>B</i>	Summe
<i>I</i>	<i>a</i>	<i>b</i>	<i>a + b</i>
<i>II</i>	<i>c</i>	<i>d</i>	<i>c + d</i>
Summe	<i>s</i>	<i>t</i>	<i>n</i>

Zur Untersuchung der Unabhängigkeit setzt Barnard ein 2-Urnen-Modell an. *A* und *B* repräsentieren diese beiden Urnen. Jede Urne enthält Kugeln der Ausprägungen *I* und *II*. Der Anteil der Kugeln mit der Markierung *I* in der Urne *A* soll mit p_a bezeichnet werden. Der Anteil der Kugeln mit Markierung *I* in *B* dementsprechend mit p_b . Barnard geht davon aus, dass die Anteile bei Unabhängigkeit beider Merkmale gleich groß sind.

³im Englischen: "Thus the new test is more powerful than Fisher's." [16]

⁴im Englischen: "That brought a reply from Fisher attacking my test. [...] Anyway, I replied to Fisher's reply and we carried on a friendly correspondence [...]" (siehe [33])

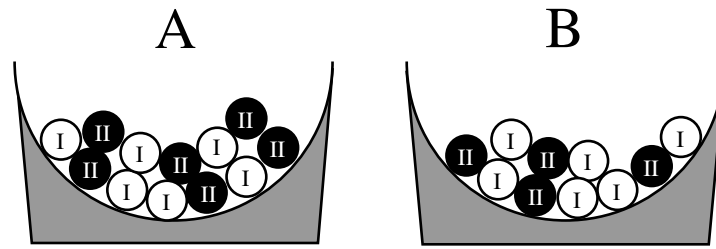


Abbildung 2.5: Modell zweier Urnen mit Zurücklegen. Anteil von I in A : p_a . Anteil von I in B : p_b .

Es lassen sich somit die Anteile und die Nullhypothese formulieren.

$$\begin{aligned}
 p_a &= P(I|A) \\
 p_b &= P(I|B) \\
 H_0 : p_a &= p_b = p
 \end{aligned} \tag{2.18}$$

Die Beobachtung der Häufigkeiten, also dem 4-Tupel (a, b, c, d) , entspricht dem Urnenmodell mit Zurücklegen. Die Anzahl der zu ziehenden Kugeln sei fest. Aus der Urne A werden s Kugeln und der Urne B t Kugeln gezogen. Die Wahrscheinlichkeit, dass dabei a Kugeln der Urne A und b Kugeln der Urne B mit Markierung I gezogen werden, lässt sich aus dem Produkt zweier binomialverteilter Größen berechnen:

$$\begin{aligned}
 P(a, b, c, d) &= \binom{s}{a} p_a^a (1 - p_a)^c \cdot \binom{t}{b} p_b^b (1 - p_b)^d \\
 &\stackrel{H_0}{=} \frac{s!t!}{a!b!c!d!} p^{a+b} (1 - p)^{c+d} =: W(a, b, p)
 \end{aligned} \tag{2.19}$$

Diese Wahrscheinlichkeit weicht um den Faktor $\frac{n!}{(a+b)!(c+d)!} p^{a+b} (1 - p)^{c+d}$ von der Berechnung von p_0 des Fisher-Yates-Testes ab (vgl. mit Formel (2.8)). Als nächstes werden die Wahrscheinlichkeiten aller Stichproben, die mindestens genauso extrem wie die Beobachtung sind, aufsummiert.

$$P(p) = \sum_{T(a', b') \geq T(a, b)} W(a, b, p) \tag{2.20}$$

$T(a, b)$ bezeichnet dabei eine geeignete Teststatistik zur Entscheidung, welches Tableau gleich extrem oder extremer ist. $W(a, b, p)$ bezeichnet die Wahrscheinlichkeit eines jeden Tableaus. Bei festen Spaltensummen hängen diese nur von a, b und dem Parameter p ab. c und d ergeben sich durch Vervollständigung der Tableaus. Daher ergibt sich die

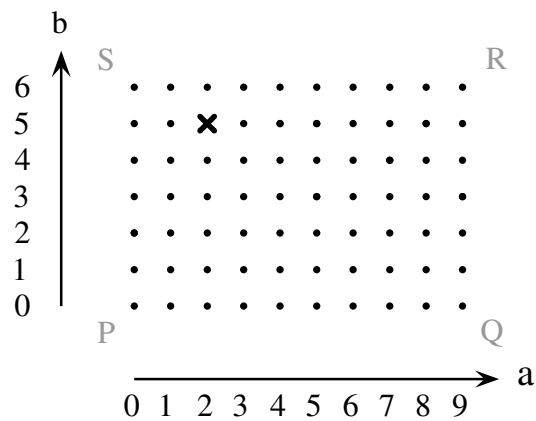


Abbildung 2.6: Lattice-Diagramm: Beispiel für $n = 15$ mit $0 \leq a \leq 9$ und $0 \leq b \leq 6$. Jeder Punkt entspricht einem Tableau. Das Kreuz entspricht der Beobachtung.

Anzahl aller möglichen Tableaus aus der Anzahl der Möglichkeiten für a und b . Diese lassen sich mit dem sogenannten Lattice-Diagramm visualisieren (siehe Abbildung 2.6). Jeder Punkt im Rechteck PQRS entspricht dabei einem Tableau. Das Kreuz bezeichnet die Beobachtung.

SCM-Bedingungen

Man überlege sich zunächst, dass Beobachtungen, die im Lattice-Diagramm auf der Diagonalen PR liegen, für gleiche Anteile $p_a = p_b$ stehen. Je weiter eine Beobachtung von dieser Diagonalen entfernt liegt, umso extremer wird die Beobachtung. Des Weiteren überlege man sich, dass Beobachtungen bzw. die Punkte (a, b) und $(s-a, t-b)$ im Lattice-Diagramm gleich extrem sein müssen, denn die Beobachtung $(s-a, t-b)$ entsteht lediglich durch Vertauschung der Bezeichnungen I und II . Die Beobachtung $(s-a, t-b)$ ist dabei für den zweiseitigen Signifikanzwert relevant. Extreme Anordnungen sind also symmetrisch zur PR-Diagonalen. Dieser Zusammenhang sei unter dem Begriff der Symmetrie-Bedingung zusammengefasst. Es genügt also zu wissen, welche Punkte (a, b) im Dreieck PRS als extrem angesehen werden können, die restlichen Punkte ergeben sich aus Spiegelung an der PR-Diagonalen - der Symmetriebedingung. Für das Dreieck PRS führen wir im mathematischen Sinne die *Größe der Abweichung* ein:

$$\begin{aligned}
 \text{Arithmetische Größe der Abweichung: } & p_a - p_b \\
 \text{Verhältnis der Abweichung: } & p_a/p_b \\
 \text{logarithmisches Verhältnis der Abweichung: } & \log(p_a/p_b)
 \end{aligned}
 \tag{2.21}$$

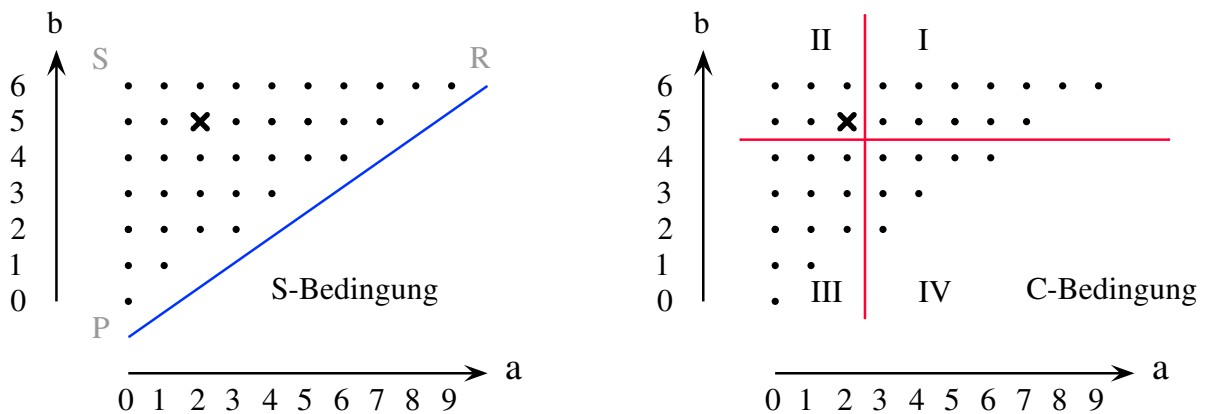


Abbildung 2.7: Symmetry-condition und convexity-condition. Extremer sind Punkte im Quadranten II, weniger extrem sind Punkte im Quadranten IV.

Wie man sich leicht erschließen kann, zeigen die Punkte direkt über und direkt links von der Beobachtung x eine größere Abweichung und sind damit als extremer zu deklarieren. Damit sind alle Punkte ausgehend von x in Richtung S extremer. Gleichzeitig impliziert diese Bedingung, dass alle Punkte in Richtung Q (bis zur Diagonalen PR) eine geringere Abweichung besitzen, demnach weniger extrem sind. Dieser Zusammenhang sei als C-Bedingung zusammengefasst⁵. GIUSEPPE CARDILLO ([51] und [52]) liefert eine Matlab-Routine zur Berechnung des Signifikanzwertes und nutzt dabei zur Entscheidung für extremere Punkte folgende Teststatistik:

$$T(a', b') = \frac{|p_a - p_b|}{\sqrt{\frac{a'+b'}{n} \frac{c'+d'}{n} (\frac{1}{s} + \frac{1}{t})}} \quad (2.22)$$

Nun ist man in der Lage die Gleichung (2.20) $P(p) = \sum_{T(a',b') \geq T(a,b)} W(a, b, p)$ aufzustellen. Unter $H_0: p_a = p_b = p$ bezeichnet $P(p)$ die Wahrscheinlichkeit der Beobachtung und mindestens gleich extremer Tableaus in Abhängigkeit von p . Das Signifikanzniveau ergibt sich aus dem Maximum von $P(p)$, $0 \leq p \leq 1$.

Matlab [Barnard-Test]

Nach Eingabe der Stichprobenvektoren x und y wird die Kontingenztafel erstellt. Dabei wird das Koordinatensystem durch die Mediane in vier Quadranten geteilt. Anschließend erfolgt die Berechnung des Signifikanz-Niveaus anhand der bestehenden Matlab-Routine `myBarnard` von GIUSEPPE CARDILLO ([51]). Die Berechnung des Niveaus erfolgt dabei

⁵C steht dabei für für das englische Wort *convexity* und bezeichnet die konvexe Anordnung extremer Punkte.

2 Von der Historie bis zur Gegenwart

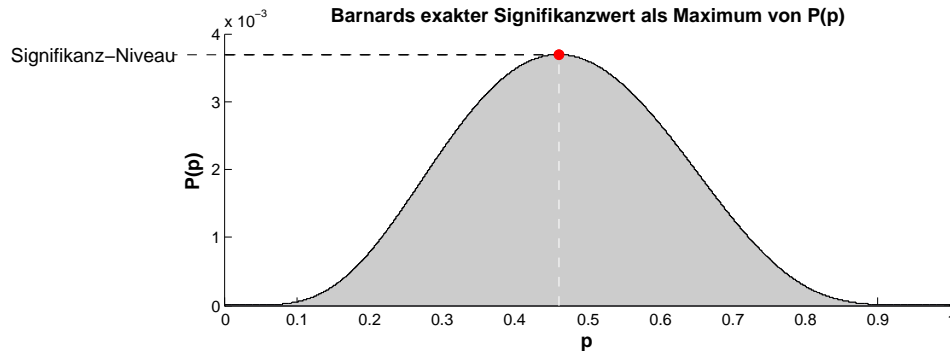


Abbildung 2.8: Beispiel zur Bestimmung des Signifikanz-Niveaus anhand des Fliegendatensatzes.

nicht exakt sondern numerisch. Der Grad der Genauigkeit kann mittels `precision` eingestellt werden – der Kehrwert von `precision` gibt die Schrittweite beim Maximierungsschritt an.

```
1 function stats=BarnardInd(x,y,precision,plot)
2 %Syntax: stats=BarnardInd(x,y,precision,plot)
3 % Calculates p-value by testing independence on the base of the
4 % implementation of Barnard's Test programmed by Cardillo G. (2009)
5 % MyBarnard: a very compact routine for Barnard's exact test on 2x2 matrix
6 % http://www.mathworks.com/matlabcentral/fileexchange/25760
7 %
8 % Marcus Vollmer
9 % 31.05.2010
10
11 S(4)=zeros;
12 n=size(x,1);
13
14 for i=1:n
15     if x(i)<median(x)
16         if y(i)>=median(y)
17             S(1)=S(1)+1;
18         else
19             S(3)=S(3)+1;
20         end
21     else if y(i)>=median(y)
22         S(2)=S(2)+1;
23     else
24         S(4)=S(4)+1;
25     end
26 end
27 end
28
29 stats=mybarnard([S(1) S(2);S(3) S(4)],precision,plot);
30 end
```


2.3.2 Hoeffdings D-Test of Independence

Wassily Hoeffding, geboren in Finnland, promoviert in Deutschland und in die Vereinigten Staaten immigriert, veröffentlichte im Jahre 1948 den Hoeffdingschen Unabhängigkeitstest⁶ [18]. Sein Anliegen war es, einen guten parameterfreien Unabhängigkeitstest zu schaffen. Unter *gut* verstand er, dass der Test erwartungstreu und konsistent ist. Ein Test sei dabei als konsistent zu bezeichnen, wenn die Wahrscheinlichkeit die Nullhypothese zu akzeptieren, für wachsende Stichprobenumfänge gegen Null geht, d.h. $P(H_0|H_1 \text{ ist wahr}) \xrightarrow{n \rightarrow \infty} 0$.

D-Test

Seien X, Y Zufallsvariablen mit stetigen Verteilungsfunktionen $F_X(x), F_Y(y)$ einer Stichprobe vom Umfang n . Die gemeinsame Verteilungsfunktion von (X, Y) sei ebenfalls stetig und mit $F_{XY}(x, y)$ bezeichnet. Im Sinne der stochastischen Unabhängigkeit beider Variablen gehen wir davon aus, dass sich die gemeinsame Verteilung aus dem Produkt der einzelnen Verteilungen berechnet: $F_{XY}(x, y) = F_X(x)F_Y(y)$. Die Differenz $D(x, y) = F_{XY}(x, y) - F_X(x)F_Y(y)$ liefert damit ein geeignetes Maß für die Abhängigkeit. Bei Unabhängigkeit erwarten wir, dass diese Differenz Null beträgt.

$$H_0 : D(x, y) = F_{XY}(x, y) - F_X(x)F_Y(y) = 0 \quad (2.23)$$

Hoeffding führt als Nächstes das Funktional $\Delta(F) = \int D^2(x, y) dF_{XY}(x, y)$ ein und stellt im Abschnitt 3 seiner Publikation [18] den Zusammenhang zu $D(x, y)$ her. Hoeffding beweist: $\Delta(F) = 0 \Leftrightarrow D(x, y) = 0$ (unter der Voraussetzung, dass $F_{XY}(x, y)$ zur Menge der Verteilungsfunktionen mit stetiger Randverteilung und stetiger gemeinsamer Verteilung gehört). Es wird weiterhin gezeigt, dass der Erwartungswert der Teststatistik D_n Δ beträgt ($\mathbb{E}(D_n) = \Delta$). Im Falle der Unabhängigkeit ist D_n um Null verteilt und eignet sich somit als Entscheidungsgröße zur Unabhängigkeit.

$$D_n = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)} \quad \text{mit} \quad (2.24)$$

⁶im Englischen: D-Test of Independence

2 Von der Historie bis zur Gegenwart

$$\begin{aligned}
 A &= \sum_{i=1}^n (\text{rg}(x_i) - 1)(\text{rg}(x_i) - 2)(\text{rg}(y_i) - 1)(\text{rg}(y_i) - 2) \\
 B &= \sum_{i=1}^n (\text{rg}(x_i) - 2)(\text{rg}(y_i) - 2)c_i \\
 C &= \sum_{i=1}^n (c_i - 1)c_i
 \end{aligned}$$

$\text{rg}(x_i)$ und $\text{rg}(y_i)$ bezeichnen dabei die Ränge von X und Y . c_i sei die Anzahl der Punkte (x_j, y_j) mit $x_j < x_i$ und $y_j < y_i$. Da $F_{XY}(x, y)$ stetig ist, gehen wir vom Fall ohne Bindungen aus, d.h. $x_i \neq x_j$ und $y_i \neq y_j$ für $i \neq j$. D_n befindet sich dann zwischen den Grenzen $-\frac{1}{60}$ und $\frac{1}{30}$, wobei die obere Schranke fest ist und die untere Schranke von nD_n gegen $-\frac{1}{36}$ läuft für $n \rightarrow \infty$.

Nach Hoeffding [18] ist D_n unter H_0 um $\mu=0$ verteilt mit einer Varianz von

$$\text{Var}(30D) = \frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)}.$$

Für $n=5, 6, 7$ gibt Hoeffding die exakte Verteilung von D_n an. Für größere Stichproben ist das Signifikanz-Niveau α nur durch eine Abschätzung gegeben, da die Art der Verteilung von D_n unbekannt ist. Hoeffding nutzt dazu die Tschebyscheffsche Ungleichung:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Durch Substitution von $\frac{1}{k^2} = \alpha$:

$$P(|X - \mu| \geq \frac{\sigma}{\sqrt{\alpha}}) \leq \alpha$$

und Ersetzen der gegebenen Größen:

$$\begin{aligned}
 P\left(30D_n \geq \sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}}\right) &\leq \\
 P\left(|30D_n| \geq \sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}}\right) &\leq \alpha
 \end{aligned}$$

gelangt man zu einer Abschätzung des Signifikanzniveaus:

$$\alpha \leq \frac{2(n^2 + 5n - 32)}{(30D_n)^2 9n(n-1)(n-3)(n-4)} \quad (2.25)$$

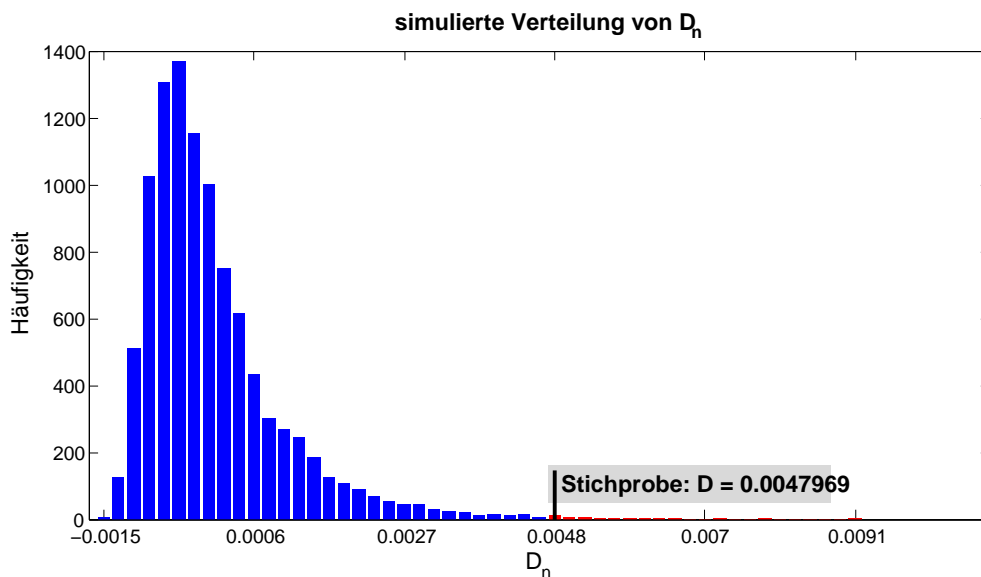


Abbildung 2.9: Beispiel für die Verteilung von D_n anhand des Fliegendatensatzes ($n=21$). Der Simulation liegen 10000 zufällige Anordnungen von n Punkten zu Grunde. Hoeffdings D-Test ergibt hierbei ein Signifikanz-Niveau von 0.0038.

Die Abschätzung von Tschebyscheff ist sichtlich schlecht und liefert keine wünschenswerte Ergebnisse. Unter Nutzung der Möglichkeiten heutiger Rechentechnik lässt sich die Verteilung von D_n unter der Nullhypothese H_0 simulieren, um eine bessere Approximation für α zu erhalten. Da die Berechnung der exakten Verteilung von D sehr aufwendig ist, entwickelten BLUM, KIEFER und ROSENBLATT [21] 1961 die einfachere Statistik $B_n = n^{-5} \sum_{i=1}^n (N_1(i)N_4(i) - N_2(i)N_3(i))^2$. $N_1(i)$, $N_2(i)$, $N_3(i)$ und $N_4(i)$ seien dabei die Anzahl der Punkte in den vier Quadranten, welche von der Horizontalen und Vertikalen des Punktes (X_i, Y_i) begrenzt werden. Dieser Ansatz soll aber in dieser Arbeit nicht weiter verfolgt werden.

Matlab [Hoeffdings D-Test]

Nach Eingabe der Stichprobenvektoren x und y wird die Teststatistik D berechnet. Ist $n > 100$, so erfolgt eine Simulation von zufälligen Anordnungen von n Punkten und die Erstellung der simulierten Verteilung von D_n . Auf dieser Basis wird der p-Wert bestimmt. Es erfolgt, falls gewünscht, die Ausgabe der Verteilung als Histogramm. Für $n \leq 100$ wird eine Quantilstabelle genutzt.

2 Von der Historie bis zur Gegenwart

```
1 function p=HoeffdingInd(x,y,plot)
2 %Syntax: p=HoeffdingInd(x,y,plot)
3 % Calculates p-value by testing independence with the Hoeffding's D-Test
4 %
5 % Inputs: x - nx1 vector of data
6 %         y - nx1 vector of data
7 %         plot - if plot=1 you will get the distribution of D
8 % Output:
9 %         p - significance level
10 %
11 % Marcus Vollmer
12 % 25.06.2010
13
14 n=size(x,1);
15 rx=tiedrank(x); ry=tiedrank(y); rxy=zeros(n,1);
16 for i=1:n
17     for j=1:n
18         if rx(j)<rx(i) && ry(j)<ry(i)
19             rxy(i)=rxy(i)+1;
20         end
21     end
22 end
23 D=teststatistic(rx,ry,rxy,n);
24
25 %Calculation of significance level
26 if n<=100
27     load('Matlab_SaveFiles/HoeffdingQuantile.mat');
28     p=(sum(quant(n-4,:)>=D))/size(quant,2);
29 else
30     %Simulation of the distribution of D = simD
31     n1=10000; n2=50;
32     simD=hoeffdingdistribution(n,n1,n2);
33     P=(sum(simD>=D))/n1;
34     p=mean(P);
35 end
36
37 %Plot
38 if plot==1 && n>100
39     figure
40     mydata=hist(simD(1:n1,1),min(simD(1:n1,1)):(max(simD(1:n1,1))-min(simD
41         (1:n1,1)))/50:max(simD(1:n1,1)));
42     bar_h=bar(mydata);
43     bar_child=get(bar_h,'Children');
44     set(bar_child,'CData',mydata);
45     mycolor=[0 0 1;1 0 0];
46
47     color=zeros(1,size(mydata,2))+2;
48     i=1; pp=mydata(i);
49     while(pp/n1<1-p)
50         color(i)=1;
51         i=i+1;
```

```

51     pp=pp+mydata(i);
52 end
53
54 set(bar_child, 'CData', color, 'EdgeColor', 'none');
55 set(gca, 'XTick', 1:10:51);
56 set(gca, 'XTickLabel', round((min(simD(1:n1,1)):(max(simD(1:n1,1))-min(
57     simD(1:n1,1)))/5:max(simD(1:n1,1))*10000)/10000);
58 set(gcf, 'PaperSize', [22 11], 'position', [100 100 800 400], 'color', [1 1
59     1]);
60 set(gca, 'OuterPosition', [0 0 1 1]);
61 title('\fontsize{14}simulated_distribution_of_D_n', 'FontWeight', 'bold')
62 ;
63 xlabel('\fontsize{14}D_n');
64 ylabel('\fontsize{14}frequency');
65 colormap(mycolor);
66 line([i i], [0 mydata(i)+max(mydata)/10], [1 1], 'LineStyle', '-', 'Color', '
67     black', 'LineWidth', 2);
68 text(i, mydata(i)+max(mydata)/10, ['\bf\fontsize{12}_sample:_D=_',
69     num2str(D)], 'HorizontalAlignment', 'left', 'VerticalAlignment', 'top', '
70     BackgroundColor', [.85 .85 .85], 'Margin', 3);
71 end
72 end
73
74 function D=hoeffdingdistribution(n,o,p)
75 D=zeros(o,p);
76 for l=1:p
77     for k=1:o
78         rx=(1:n)';
79         ry=randperm(n)';
80         rxy=zeros(n,1);
81         for i=1:n
82             for j=1:i
83                 if ry(j)<ry(i)
84                     rxy(i)=rxy(i)+1;
85                 end
86             end
87         end
88         D(k,l)=teststatistic(rx,ry,rxy,n);
89     end
90 end
91
92 function D=teststatistic(rx,ry,rxy,n)
93 A=sum((rx-1).*(rx-2).*(ry-1).*(ry-2));
94 B=sum((rx-2).*(ry-2).*rxy);
95 C=sum((rxy-1).*rxy);
96 D=(A-2*(n-2)*B+(n-2)*(n-3)*C)/(n*(n-1)*(n-2)*(n-3)*(n-4));
97 end

```

2.3.3 Bakirovs I_n

BAKIROV, RIZZO und SZÉKELY entwickelten 2004 einen neuen parameterfreien Test zur Unabhängigkeit. Konzipiert für das Testen der Unabhängigkeit zweier oder mehrerer Vektoren beliebiger Dimensionen, wurde dieser anschließend im Jahre 2006 veröffentlicht [41]. Die Teststatistik basiert dabei auf Distanzen zweier Punkte und hat dabei sehr gute Eigenschaften, die wir im Weiteren kennen lernen werden. Doch beginnen wir zuerst mit den gegebenen Größen für den bivariaten Fall.

Gegeben seien zwei Vektoren X und Y der reellen Zahlen. Bezeichnet man die Verteilungsfunktionen von X und Y mit F_1 und F_2 , sowie die gemeinsame Verteilungsfunktion mit F , so lässt sich für unabhängige Vektoren X und Y die Nullhypothese aufstellen:

$$H_0 : F = F_1 F_2$$

Mit Hilfe einer zufälligen Stichprobe Z vom Umfang n mit $Z_j=(X_j, Y_j)$, $X_j \in \mathbb{R}$, $Y_j \in \mathbb{R}$, $j=1, \dots, n$ wollen wir die Hypothese testen. Folgende Teststatistik wird nach BAKIROV ET AL. als sinnvoll erachtet. Mit $|\cdot|_d$ als Euklidische Norm des \mathbb{R}^d und $Z_{ij}=(X_i, Y_j)$ lautet diese:

$$I_n = \sqrt{\frac{2\bar{z} - z_2 - z}{x + y + z}} \quad (2.26)$$

wobei

$$\begin{aligned} z_2 &= \frac{1}{n^2} \sum_{i,j=1}^n |Z_{ii} - Z_{jj}|_2 & x &= \frac{1}{n^2} \sum_{i,j=1}^n |X_i - X_j|_1 \\ \bar{z} &= \frac{1}{n^3} \sum_{i,j=1}^n \sum_{k=1}^n |Z_{ii} - Z_{jk}|_2 & y &= \frac{1}{n^2} \sum_{i,j=1}^n |Y_i - Y_j|_1 \\ z &= \frac{1}{n^4} \sum_{i,j=1}^n \sum_{k,l=1}^n |Z_{ij} - Z_{kl}|_2 \end{aligned}$$

Begründet wird die Wahl dieser Teststatistik mit der normierten Distanz zwischen der gemeinsamen charakteristischen Funktion $f(t, s)$ von (X, Y) und dem Produkt der marginalen charakteristischen Funktionen $f_1(t)$ und $f_2(s)$ von X und Y :

$$I = I(f) = \frac{\|f(s, t) - f_1(t)f_2(s)\|}{\|\sqrt{(1 - |f_1(t)|^2)(1 - |f_2(s)|^2)}\|} \quad (2.27)$$

I ist ein Maß für die Abhängigkeit und I_n ergibt sich als empirisches Gegenstück zu I . Details sind der Veröffentlichung „*A multivariate nonparametric test of independence*“ [41] zu entnehmen. Es konnte von BAKIROV ET AL. gezeigt werden, dass I_n invariant

gegenüber Verschiebungen, orthogonalen Transformationen und gemeinsamen Skalierungen von X und Y ist. Als vorteilhaft wird auch bemerkt, dass für die Berechnung keine Matrixinversion gebraucht wird. Des weiteren werden folgende Eigenschaften für I und I_n getroffen:

- (i) $\mathbb{E}(|X|_1 + |Y|_1) < \infty \Rightarrow I_n \xrightarrow[n \rightarrow \infty]{\text{f. s.}} I$ ⁷.
- (ii) $0 \leq I_n \leq 1$.
- (iii) $I = 0 \Leftrightarrow X$ und Y sind Unabhängig.
- (iv) $I = 1 \Leftrightarrow \exists$ Zufallsmenge A und Vektoren a, b, c, d , sodass $X = a + b\chi(A)$ und $Y = c + d\chi(A)$, wobei $\chi(A)$ die Indikatorfunktion der Zufallsmenge A beschreibt.
- (v) Bei Unabhängigkeit von X und Y , sowie Bedingung (i): $nI_n^2 \xrightarrow[n \rightarrow \infty]{D} Q$ ⁸.
 Q bezeichnet eine nichtnegative quadratische Form normalverteilter Zufallsvariablen mit $\mathbb{E}(Q) = 1$.

Die wichtigste Erkenntnis zur Bestimmung des Signifikanz-Niveaus dürfte Bedingung (v) sein. Denn für Q gilt folgende Ungleichung, wenn $0 < \alpha \leq 0.215$:

$$\mathbb{P}(Q \geq (\Phi^{-1}(1 - \frac{\alpha}{2}))^2) \leq \alpha \quad (2.28)$$

und aus (v) folgt, dass $\forall \alpha \in]0; 0.215]$ H_0 abgelehnt wird, falls:

$$\sqrt{n}I_n \geq \Phi^{-1}(1 - \frac{\alpha}{2}). \quad (2.29)$$

Damit gelangt man durch Umformungen zu:

$$\alpha < 2(1 - \Phi(\sqrt{n}I_n)). \quad (2.30)$$

Weil dies aber nur in der Konvergenz der Fall ist, sollten wir für kleine Stichprobengrößen auf eine Simulation der Verteilung von I zurückgreifen.

⁷Mit f.s. ist der Begriff der *fast sicheren* Konvergenz verbunden, d.h. $\mathbb{P}(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}) = 1$

⁸D steht dabei für *convergence in distribution* (schwache Konvergenz), d.h. $F(nI_n^2)$ konvergiert gegen $F(Q)$ für $n \rightarrow \infty$.

Matlab [Bakirovs Independence Test]

Mit Eingabe der Stichprobenvektoren x und y erfolgt die Berechnung der Teststatistik I_n , sowie die Berechnung des Signifikanzniveaus nach (2.30).

```
1 function p=BakirovInd(x,y)
2 %Syntax: p=BakirovInd(x,y)
3 % Calculates p-value by testing independence with the test of independence
4 % by Bakirov/Rizzo/Székely (2006)
5 %
6 % Inputs: x - nx1 vector of data
7 %         y - nx1 vector of data
8 % Output: p - significance level
9 %
10 % Marcus Vollmer
11 % 08.07.2010/11.07.2010
12
13 n=size(x,1);
14 I_n=teststatistic(x,y,n);
15
16 if n<101
17     load('Matlab_SaveFiles/BakirovQuantile.mat','quant');
18     p=(sum(quant(n,1:size(quant,2))>=I_n)+1)/(size(quant,2)+1);
19 else
20     p=2*(1-normcdf(sqrt(n)*I_n));
21     if p>0.215
22         disp('Warning: p-value could be wrong. ')
23     end
24 end
25 end
26
27 function I_n=teststatistic(x,y,n)
28 z1=0; zquer=0;
29 diffx=(repmat(x,1,n)-repmat(x',n,1));
30 diffy=(repmat(y,1,n)-repmat(y',n,1));
31 z2=sum(sum(sqrt(diffx.^2+diffy.^2)))/n^2;
32 x1=sum(sum(sqrt(diffx.^2)))/n^2;
33 y1=sum(sum(sqrt(diffy.^2)))/n^2;
34
35 for i=1:n
36     for j=1:n
37         z1=z1+sum(sum(sqrt(diffx.^2+diffy(i,j).^2)));
38     end
39     zquer=zquer+sum(sum(sqrt(diffx.^2+repmat(diffy(:,i),1,n).^2)));
40 end
41 zquer=zquer/n^3;
42 z1=z1/n^4;
43
44 %teststatistic I_n
45 I_n=sqrt((2*zquer-z2-z1)/(x1+y1-z1));
46 end
```


2.3.4 Permutation-Entropy-Test von García-Marín

Der Permutation-Entropy-Test von M. MATILLA-GARCÍA und M. RUIZ MARÍN [44] ist ein parameterfreier und konsistenter Unabhängigkeitstest. Eigentlich ist dieser Test als Test zur Messung der Abhängigkeit für Zeitreihen konzipiert. Wenn wir jedoch zwei stetige Variablen X und Y haben, deren Ränge bilden und die Werte von Y nach X sortieren, so können wir die sortierten Ränge von Y als Zeitreihe auffassen. Der Permutation-Entropy-Test benutzt dabei die Verteilung ordinaler Muster in der Zeitreihe. Als ein ordinales Muster bezeichnen wir eine Ordnungsrelation m benachbarter Werte und geht auf C. BANDT und B. POMPE [39] zurück, welche auch die Definition der Permutation-Entropy liefern. Beginnen wir zunächst mit den Definitionen und Bezeichnungen.

X und Y seien n -dimensionale stetige Variablen der reellen Zahlen. $\text{rg}(X)$ bezeichnet den Rang von X . Die Paare $((X_i), (Y_i))_{i=1, \dots, n}$ werden nun nach den Rängen von X geordnet⁹: $(\text{rg}(X_j)=i, (Y_j))_{i=1, \dots, n}$. Die Folge der (Y_j) ist als Zeitreihe anzusehen und sei mit Z bezeichnet. Als nächstes betrachten wir die ordinalen Muster von m benachbarten Werten: $Z_m(t) = (Z(t), Z(t+1), \dots, Z(t+m-1))$ mit $t=1, \dots, n-m+1$. m wird als Bindungsdimension (embedding dimension) bezeichnet und ist eine positive ganze Zahl größer 1, also $m \geq 2$; je nachdem, wieviele benachbarte Werte verglichen werden sollen. Mögliche ordinale Muster sind Elemente der symmetrischen Gruppe S_m , also der Gruppe, die aus allen Permutationen der Menge $(0, \dots, m-1)$ besteht. So sind für $m = 3$ folgende ordinale Muster möglich: $S_3 = \{(0, 1, 2), (0, 2, 1), (1, 0, 2), (1, 2, 0), (2, 0, 1), (2, 1, 0)\}$. Wir weisen jedem $(Z_m(t))_{t=1, \dots, n-m+1}$ genau jenes Muster $\pi_i = (i_1, i_2, \dots, i_m) \in S_m$ zu, für welches gilt:

$$Z_{t+i_1} \leq Z_{t+i_2} \leq \dots \leq Z_{t+i_m} \quad \text{und} \quad (2.31)$$

$$i_s < i_{s+1} \quad \text{für} \quad Z_{t+i_s} = Z_{t+i_{s+1}} \quad (2.32)$$

Die beiden Bedingungen garantieren, dass jedes $(Z_m(t))_{t=1, \dots, n-m+1}$ genau einem Muster zugeordnet wird. Beginnen wir mit einem kleinen Beispiel.

Gegeben sei eine endliche Zeitreihe mit 8 Werten: $Z = \{6, 7, 2, 7, 5, 1, 3, 4\}$. Wir analysieren die Zeitreihe bei einer Bindungsdimension von $m=3$. Für $t=1$ vergleichen wir nun die ersten 3 Werte und nach Bedingung (2.31) ergibt sich eine Ordnung: $Z_{t+2}=2 < Z_{t+0}=6 < Z_{t+1}=7$ und somit ist $Z_3(t=1)$ dem ordinalen Muster $(2, 0, 1)$ zuzuordnen. Für $t=2$ gilt:

⁹Dabei sei der Fall von Bindungen, also der Zuweisung gleicher Ränge für identische Werte ausgeschlossen. Eventuell muss vorab ein „verzittern“ der Werte von X vorgenommen werden. Dies kann beispielsweise durch Addition normalverteilter Werte mit Mittelwert 0 und einer kleinen Varianz geschehen. Dadurch werden die Bindungen aufgelöst.

2 Von der Historie bis zur Gegenwart

$Z_{t+1}=2 < Z_{t+0}=Z_{t+2}=7$. Da $Z_{t+0}=Z_{t+2}$, wird $Z_3(t=2)$, aufgrund von Bedingung (2.32), dem ordinalen Muster $(1, 0, 2)$ zugeordnet. Dies werde nun schrittweise fortgeführt. Es zeigt sich, dass in unserem Beispiel jedes ordinale Muster genau ein mal vorkommt:

ordinales Muster						
symmetrische Gruppe	0 1 2	0 2 1	1 0 2	1 2 0	2 0 1	2 1 0
zugeordnete Zeitreihenausschnitte	1, 3, 4 $Z_3(6)$	2, 7, 5 $Z_3(3)$	7, 2, 7 $Z_3(2)$	5, 1, 3 $Z_3(5)$	6, 7, 2 $Z_3(1)$	7, 5, 1 $Z_3(4)$

Abbildung 2.10: Ordinale Muster der Bindungsdimension $m=3$ und deren zugeordnete Zeitreihenausschnitte des Beispiels.

Dass die ordinalen Muster zu gleichen Anteilen in der Zeitreihe auftreten, würden wir für unabhängige Zeitreihen erwarten. Dieser Fakt erlaubt es, die Permutation-Entropie $h(m)$ in die Teststatistik einfließen zu lassen:

$$h(m) = - \sum_{i=1, \dots, m!} p_{\pi_i} \ln(p_{\pi_i}) \quad (2.33)$$

mit den relativen Häufigkeiten p_{π_i} der ordinalen Muster π_i :

$$p_{\pi_i} = \frac{n_{\pi_i}}{n - m + 1} := \frac{\#\{t | Z_m(t) \text{ hat Muster } \pi\}}{n - m + 1} \quad (2.34)$$

Bei monotonen Zusammenhängen ergibt sich für die Entropie eine untere Grenze von 0. Sind alle Muster gleich häufig vertreten, ergibt sich die obere Grenze $\ln(m!)$ und damit:

$$0 \leq h(m) \leq \ln(m!)$$

Um zu einer geeigneten Teststatistik zu gelangen, betrachten wir die Zeitreihe Z . Wir definieren nun die Zufallsvariable $\tilde{X}_{\pi_i t}$ über alle möglichen ordinalen Muster $\pi_i = (i_1, i_2, \dots, i_m) \in S_m$:

$$\tilde{X}_{\pi_i t} = \begin{cases} 1 & \text{falls } Z_{t+i_1} \leq Z_{t+i_2} \leq \dots \leq Z_{t+i_m}, \\ 0 & \text{sonst} \end{cases} \quad (2.35)$$

Wir können $\tilde{X}_{\pi_i t}$ als Bernoulli-Variable betrachten, mit Erfolgswahrscheinlichkeit p_{π_i} für das ordinale Muster i . Die Summe dieser Wahrscheinlichkeiten ist 1. Ist unsere Zeitreihe endlich und hat n Werte, so lässt sich für jedes Muster die Variable

$$\tilde{Y}_{\pi_i} = \sum_{t=1}^{n-m+1} \tilde{X}_{\pi_i t}$$

definieren. Die Variable \tilde{Y}_{π_i} kann nur Werte zwischen 0 und $n-m+1$ annehmen. Es folgt, dass \tilde{Y}_{π_i} binomialverteilt ist. Unter der Voraussetzung der Unabhängigkeit können wir für Anzahlen $a_1, \dots, a_{m!} \in \mathbb{N} \cup \{0\}$, mit $\sum a_i = n-m+1$, die folgende Wahrscheinlichkeit bilden:

$$P(Y_{\pi_1}=a_1, Y_{\pi_2}=a_2, \dots, Y_{\pi_{m!}}=a_{m!}) = n! \prod_{i=1}^{m!} \frac{p_{\pi_i}^{a_i}}{a_i!} \quad (2.36)$$

Das entspricht einer Multinomialverteilung. Wir nutzen die Log-Likelihood-Methode um die wahren Anteile p_{π} zu schätzen. Dies ergibt den Likelihoodschätzer $\hat{p}_{\pi_i} = \frac{n_{\pi_i}}{n-m+1}$ für die Anteile p_{π_i} der i ordinalen Muster. Als nächstes bilden wir den Likelihood-Quotienten:

$$\lambda(\tilde{Y}) = \frac{n! \prod_{i=1}^{m!} \frac{p_{\pi_i}^{n_{\pi_i}}}{n_{\pi_i}!}}{n! \prod_{i=1}^{m!} \frac{\binom{n_{\pi_i}}{n}^{n_{\pi_i}}}{n_{\pi_i}!}} = n^n \prod_{i=1}^{m!} \left(\frac{p_{\pi_i}}{n_{\pi_i}} \right)^{n_{\pi_i}} \quad (2.37)$$

mit dem Prüfwert:

$$G(m) = -2 \ln(\lambda(\tilde{Y})) \stackrel{p_{\pi_i} = \frac{1}{m!}}{=} 2(n-m+1)(\ln(m!) - h(m)) \quad (2.38)$$

Die Details der Umformungen von (2.38) ist der Original-Veröffentlichung [44] zu entnehmen. Bekanntermaßen ist $G(m)$ asymptotisch χ^2 -verteilt mit $m!-1$ Freiheitsgraden. Für die Teststatistik $G(m)$ lässt sich nun mehr bei gegebenen α -Niveau ein Ablehnungsbereich angeben:

$$\begin{aligned} H_0 : \quad & \{Y_1, \dots, Y_n\} \text{ i.i.d.} \\ \text{Lehne } H_0 \text{ ab, wenn } & G(m) \geq \chi_{\alpha}^2. \end{aligned}$$

Um die χ^2 -Verteilung auch anwenden zu können, sollte darauf geachtet werden, dass die erwartete Häufigkeit eines ordinalen Musters ≥ 5 ist. Demnach sollten wir die χ^2 -Verteilung erst ab einer Stichprobengröße von $5m!$ verwenden. Für eine Bindungsdimension von $m=3$ also erst ab Stichprobenumfängen ≥ 30 .

Matlab [Permutation-Entropy-Test von García-Marín]

Nach Eingabe der Stichprobenvektoren x , y und der Bindungsdimension m erfolgt die Berechnung der Teststatistik G , sowie die Berechnung des Signifikanzniveaus unter Nutzung einer Quantilstabelle oder für größere Stichprobenumfänge durch Nutzung der asymptotischen χ^2 -Verteilung.

```

1  function [p pmin pmax]=GarciaMarinInd(x,y,m)
2  %Syntax: p=garciamarinind(x,y,m)
3  % Calculates p-value by testing independence with the permutation entropy
4  % test by M. Martilla-Garcia and M. Ruiz Marin (2008)
5  %
6  % Inputs:      x - nx1 vector of data
7  %              y - nx1 vector of data
8  %              m - positive integer, embedding dimension
9  %
10 % Outputs:     p - significance level
11 %              pmax - upper bound of p
12 %              pmin - lower bound of p
13 %
14 % Marcus Vollmer
15 % 13.07.2010-15.07.2010
16
17 n=size(x,1);
18
19 if m<3
20     m=3; disp('Embedding_dimension_must_be_3_at_least , m_is_set_to_3. ')
21 else if n<factorial(m)
22     m=3; disp('Sample_size_too_small_for_your_embedding_dimension , m_is_set
23             _to_3. ')
24 end
25
26 %order pairs (x,y) and generate ranks
27 [o TIEDADJ]=tiedrank(sortrows([x y]));
28 if TIEDADJ(1)~=0 %jitter x points if ranks are tied
29     x=x+normrnd(0,(max(x)-min(x))/n/1000,n,1);
30     o=tiedrank(sortrows([x y]));
31 end
32 if TIEDADJ(2)~=0 %condition 2 if ranks are tied
33     o=tiedrank(sortrows(sortrows([x tiedrank(y)],[2 1])+[zeros(1,n);1:n]'))
34     ;
35 end
36
37 z=zeros(m,n-m+1);
38 prodvec=zeros(m,n-m+1);
39 temp=0;
40 for i=1:m
41     z(i,1:n-m+1)=o(i:n-m+i,2);
42     prodvec(i,1:n-m+1)=(m)^(i-1);
43     temp=temp+(i-1)*m^(i-1);

```

2.3.4 Permutation-Entropy-Test von García-Marín

```

43 end
44 zo=tiedrank(z);
45 no=sum((zo-1).*prodvec);
46
47 counts=zeros(temp,1);
48 for i=1:n-m+1
49     counts(no(i))=counts(no(i))+1;
50 end
51
52 ppi=counts./(n-m+1);           %frequencies of symmetric groups
53 h=-1*nansum(ppi.*log(ppi));   %Shannon entropy
54 G=2*(n-m+1)*(log(factorial(m))-h); %Teststatistic G
55
56 if n>100
57     p=1-chi2cdf(G, factorial(m)-1); %G is asymptotically chi^2 distributed
58     pmin=p;
59     pmax=p;
60 else                               %Using tables of quantiles
61     switch m
62         case 3;
63             load('Matlab_SaveFiles/GarciaMarinQuantile.mat','quant3');
64             load('Matlab_SaveFiles/GarciaMarinQuantile.mat','quantvar3');
65             quant=quant3; quantvar=quantvar3;
66         case 4;
67             load('Matlab_SaveFiles/GarciaMarinQuantile.mat','quant4');
68             load('Matlab_SaveFiles/GarciaMarinQuantile.mat','quantvar4');
69             quant=quant4; quantvar=quantvar4;
70     end
71     l=n+1-factorial(m);
72     p=(sum(quant(1,:)>=G)+1)/(size(quant,2)+1);
73     pmin=p;
74     pmax=p;
75     %Calculation of the confidence interval of p
76     if p~=0
77         alpha=0.05;
78         var=quantvar(max(sum(quant(1,:)>=G),1));
79         Gmin=G-sqrt(20*var/19)*tinv(1-alpha/2,19);
80         Gmax=G+sqrt(20*var/19)*tinv(1-alpha/2,19);
81         pmin=(sum(quant(1,:)>=Gmax)+1)/(size(quant,2)+1);
82         pmax=(sum(quant(1,:)>=Gmin)+1)/(size(quant,2)+1);
83     end
84 end
85 end

```

2.3.5 Längste aufsteigende Teilfolge L_n

Ein sehr junger Ansatz auf Unabhängigkeit zu testen stammt aus dem letzten Jahr und wurde von JESÚS E. GARCIA und VERÓNICA A. GONZÁLEZ-LÓPEZ veröffentlicht [45]. Ihre Idee beruht auf der Länge L_n der längsten aufsteigenden Teilfolge (LIS¹⁰) einer Punktmenge mit n Koordinaten¹¹ und ist parameterfrei. So ist die Länge der LIS bei monotonen Zusammenhängen entweder 1 für negative monotone Zusammenhänge oder genauso groß wie der Stichprobenumfang n bei positiven monotonen Zusammenhängen. Die Unabhängigkeit soll sich also mit dem Argument widerlegen lassen, dass die Länge der LIS entweder zu groß oder zu gering ist. Haben wir die Verteilung von L_n für unabhängige Variablen, so können wir auf Basis dieser Verteilung eine Entscheidung fällen. Ob und wie diese Verteilung exakt berechnet werden kann, werden wir auf den folgenden Seiten ergründen. Doch beginnen wir, wie üblich, mit den Definitionen und Bezeichnungen.

X und Y seien zwei stetige Zufallsvariablen. Wir testen die Nullhypothese $H_0 : X$ und Y sind unabhängig. Zur Überprüfung dessen nehmen wir eine Stichprobe (X, Y) der Größe n und bezeichnen die Paare mit $(x_1, y_1), \dots, (x_n, y_n)$. $rg(x_i)$ und $rg(y_i)$ bezeichnen die Ränge des Paares (x_i, y_i) . Wir sortieren die Paare aufsteigend nach X . Bilden wir die Ränge der entstanden Folge, so erhalten wir die Paare $\{(1, i_1), (2, i_2), \dots, (n, i_n)\}$ mit $i_j \in \{1, \dots, n\}$, für paarweise verschiedene Werte ($i_j \neq i_k$ für $j \neq k$). Die Menge dieser Paare entspricht einer Permutation π_s mit: $\pi_s(1)=i_1, \pi_s(2)=i_2, \dots, \pi_s(n)=i_n$. Die Information der aufsteigenden und absteigenden Folgen ist nun in der Folge $\{i_j\}_{j=1}^n$ enthalten.

Aufsteigende Teilfolge (IS)

S_n bezeichnet die Permutationsgruppe der Menge $\{1, \dots, n\}$. Für eine Permutation $\pi \in S_n$ bezeichnen wir eine Folge $\pi(j_1), \dots, \pi(j_k)$ als eine aufsteigende Teilfolge von π der Länge k , wenn

$$1 \leq j_1 \leq j_2 \leq \dots \leq j_k \leq n \quad \text{und} \quad (2.39)$$

$$1 \leq \pi(i_1) < \pi(i_2) < \dots < \pi(i_k) \leq n \quad (2.40)$$

Längste aufsteigende Teilfolge (LIS)

Wir bezeichnen eine aufsteigende Teilfolge $\pi(j_1), \dots, \pi(j_k)$ als längste aufsteigende Teilfolge, wenn keine aufsteigende Teilfolge $\pi(j_1), \dots, \pi(j_{\hat{k}})$ existiert mit $\hat{k} > k$. Die Länge der

¹⁰LIS steht dabei für *Longest Increasing Subsequence*.

¹¹Genauso gut können wir auch die längste absteigende Teilfolge nehmen oder auch beide gleichzeitig betrachten.

längsten aufsteigenden Teilfolge sei mit $L_n(\pi)$ bezeichnet¹². Analog lässt sich die längste absteigende Teilfolge definieren. Deren Länge sei mit $LD_n(\pi)$ bezeichnet. Mit p_k^n bezeich-

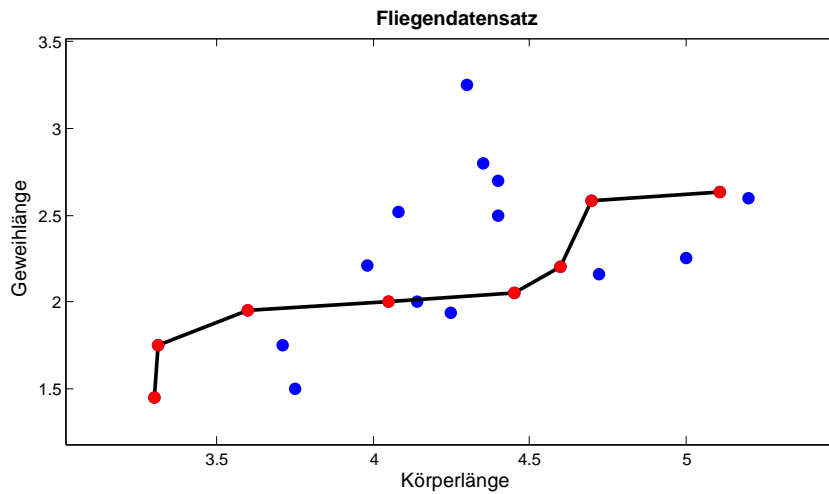


Abbildung 2.11: Eine längste aufsteigende Teilfolge des Fliegendatensatzes mit 21 Datenpaaren. Die Länge der LIS beträgt 8.

nen wir die Wahrscheinlichkeit, dass eine zufällige Permutation π der Länge n mit $\pi \in S_n$ eine längste aufsteigende Teilfolge der Länge k hat:

$$p_k^n := P(L_n=k) = \frac{\#\{\pi \in S_n : L_n(\pi)=k\}}{n!} \tag{2.41}$$

Doch **wie viele Permutationen haben eine längste aufsteigende Teilfolge der Länge k ?** Zum Glück waren solche Folgen bereits Gegenstand umfassender Forschung. So lieferte C. SCHENSTED bereits 1961 die Antwort auf diese Frage (siehe [22]):

Satz von Schensted

Die Anzahl der Permutationen n paarweise verschiedener Zahlen, die eine längste aufsteigende Teilfolge der Länge k und eine längste absteigende Teilfolge der Länge l haben, ist die Summe aus der quadrierten Anzahl der Standard-Young-Tableaus (SYT), deren Typen k Spalten und l Zeilen haben:

$$\#\{\pi \in S_n : L_n(\pi)=k \text{ und } LD_n(\pi)=l\} = \sum_{\text{Typen}} \#^2\{\text{SYT des Typs mit } k \text{ Spalten } l \text{ Zeilen}\} \tag{2.42}$$

¹²Die Definition der LIS ist nicht eindeutig. Es können mehrere längste aufsteigende Teilfolgen existieren. Vordergründig ist allerdings die Länge, nicht die Folge selbst.

Summieren wir über die Länge der längsten absteigenden Teilfolge LD_n , so erhalten wir die gewünschte Anzahl. Zum Verständnis sei im Folgenden das Standard-Young-Tableau und deren Typ definiert.

Standard-Young-Tableau (SYT)

Ein Standard-Young-Tableau der Ordnung n ist eine Anordnung von n paarweise verschiedenen natürlichen Zahlen in Zeilen und Spalten, so dass die Zahlen in jeder Zeile und Spalte von links nach rechts, bzw. von oben nach unten lückenlos und aufsteigend geordnet sind und die erste Spalte jeweils die erste Zahl einer Zeile angibt¹³.

Lassen wir die Bedingung der aufsteigenden Ordnung in Zeilen und Spalten weg, so erhalten wir die Definition des Young-Tableaus. Bevor wir uns mit der Anzahl der SYTs beschäftigen, klären wir den Typ eines SYTs. Für die Ordnung $n=4$ können wir leicht visualisieren, wie die Standard-Young-Tableaus aussehen können:

Typ eines Standard-Young-Tableaus

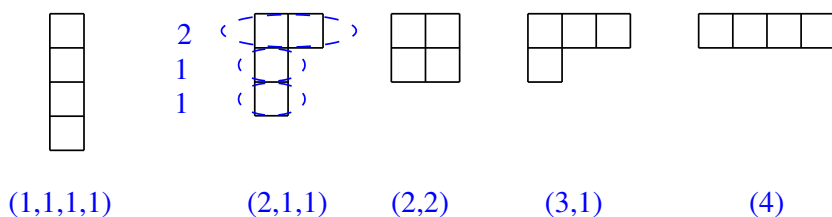


Abbildung 2.12: Mögliche Typen der Standard-Young-Tableaus der Ordnung 4.

Typ eines Standard-Young-Tableaus

Als Typ eines Standard-Young-Tableaus bezeichnen wir die optische Gestalt eines SYTs, welche aus der Anzahl der Elemente jeder Zeile definiert wird. Hat ein Tableau l Zeilen mit jeweils k_1, \dots, k_l Elementen, so ist der Typ dieses Tableaus mit λ zu bezeichnen und $\lambda=(k_1, \dots, k_l)$.

Jeder Typ eines SYTs kann mit paarweise verschiedenen natürlichen Zahlen so belegt werden, dass die Eigenschaften der Definition des SYT's nicht verletzt werden. Die Anzahl der Möglichkeiten für einen Typ λ sei mit f_λ bezeichnet. Abbildung 2.13 zeigt die möglichen SYTs des Typs $(2, 2, 1)$ und $f_{2,2,1} = 5$. Über die allgemeine Anzahl der Möglichkeiten der

¹³Diese Eigenschaft könnten wir auch mit *linksbündig* bezeichnen.

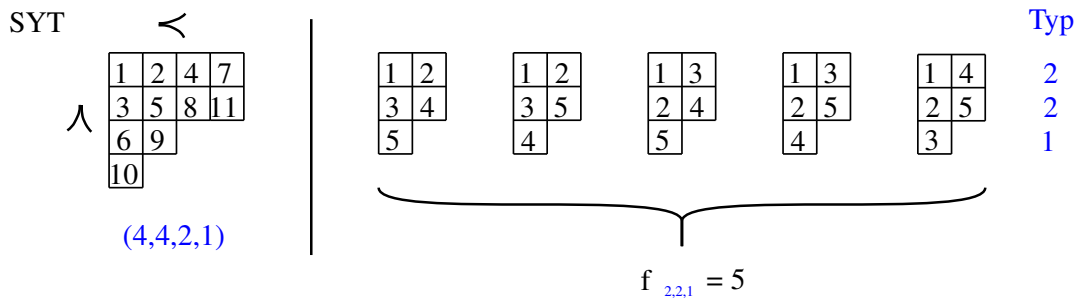


Abbildung 2.13: Links: Standard-Young-Tableau des Typs $\lambda = (4, 4, 2, 1)$ ist lexikographisch geordnet. Rechts: Alle möglichen SYTs des Typs $(2, 2, 1)$.

Vervollständigung eines SYTs von Typ λ hilft uns der Satz von FRAME, ROBINSON und TRALL [20]:

Satz von Frame-Robinson-Thrall

Die Anzahl an Standard-Young-Tableaus der Ordnung n und des Typ λ ist:

$$f_\lambda = \frac{n!}{\prod_{j=1}^n h_j} \tag{2.43}$$

Wobei h_j die Hookzahl des j -ten Elements der SYTs bezeichnet. Die Hookzahl, oder auch Hooklänge, ist die Anzahl der Elemente, die rechts und unter dem Element j liegen, addiert mit 1.

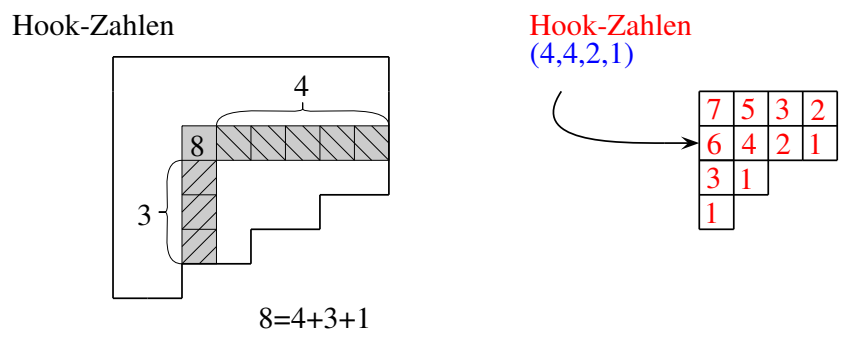


Abbildung 2.14: Links: Berechnung der Hookzahl eines Elementes des SYT aus der Anzahl der Elemente, die rechts und darunter liegen plus 1 (dem Element selbst). Rechts: Beispiel der Hookzahlen für den Typ $\lambda = (4, 4, 2, 1)$.

Beispielsweise ist $f_{4,4,2,1} = \frac{11!}{7 \cdot 5 \cdot 3 \cdot 2 \cdot 6 \cdot 3 \cdot 2 \cdot 1 \cdot 3 \cdot 1 \cdot 1} = 1760$

2 Von der Historie bis zur Gegenwart

Fassen wir nun die Ergebnisse aus dem Satz von Schensted (2.42), dem Satz von Frame et al. (2.43) und der Berechnungsvorschrift (2.41) in einer Formel zusammen:

$$p_k^n = \frac{1}{n!} \sum_{l=1}^n \sum_{\lambda \in V(k,l)} (f_\lambda)^2 \quad (2.44)$$

$V(k, l)$ bezeichnet dabei die Menge der Typen von Standard-Young-Tableaus der Ordnung n , die k Spalten und l Zeilen besitzen.

Wir sind nun in der Lage, die exakte Verteilung der Längen L_n anzugeben. Hat unsere Stichprobe der Größe n eine längste aufsteigende Teilfolge der Länge l_0 , so berechnet sich der zweiseitige P-Wert p aus:

$$p = \min \left\{ 1, 2 \cdot \sum_{k=1}^{l_0} p_k^n, 2 \cdot \sum_{k=l_0}^n p_k^n \right\} \quad (2.45)$$

Können wir noch für kleine n die exakte Verteilung berechnen, empfiehlt es sich, für größere Umfänge, auf die asymptotische Verteilung der L_n zurückzugreifen.

Asymptotische Verteilung von L_n

BAIK ET AL. [37] haben Folgendes bewiesen:

Sei S_n die Gruppe aller gleichverteilten Permutationen von n Zahlen und sei $L_n(\pi)$ die Länge der längsten ansteigenden Teilfolge von $\pi \in S_n$. χ sei eine Zufallsvariable mit Verteilungsfunktion F_{TW} (Tracy-Widom-Verteilung). Dann konvergiert χ_n schwach gegen χ für $n \rightarrow \infty$.

$$\chi_n = \frac{L_n - 2\sqrt{n}}{n^{1/6}} \xrightarrow[n \rightarrow \infty]{D} \chi \quad (2.46)$$

Für weitere Informationen zur Tracy-Widom-Verteilung F_{TW} sei auf BAIK ET AL. [37] verwiesen. Zur Berechnung des p -Wertes nutzen wir die Quantile dieser Verteilung. Die entsprechenden Quantile einiger Testgrößen finden sich in einer Tabelle im Anhang der Masterarbeit von A. BEJAN [42].

Matlab [Längste aufsteigenden Teilfolge]

Mit Eingabe der Stichprobenvektoren x, y erfolgt die Berechnung der Länge der längsten aufsteigenden Teilfolge. Für Stichprobengrößen ≤ 100 berechnet sich das Signifikanz-Niveau nach der exakten Verteilung, welche vorab berechnet wurde. Ich nutzte dafür den ZS1-Algorithmus von ZOGHBI und STOJMENOVIC [36] und zur Vorallozierung des

Speichers die Partitionsfunktion von Numerica.com [49]. Die dadurch erstellten Partitionen sind Typen des Standard-Young-Tableaus. Durch Berechnung der zugehörigen Hookzahlen lässt sich mittels (2.44) die Verteilung der Längsten aufsteigenden Teilfolge exakt angeben. Die aufwendige Berechnung der Partitionen erlaubt eine exakte Angabe für Stichprobengrößen von 1 bis 100. Für größere Stichprobenumfänge wird die asymptotischen Tracy-Widom-Verteilung genutzt.

```

1 function [p seq]=LisTestInd(x,y,plot)
2 %Syntax: [p seq]=LisTestInd(x,y,plot)
3 % Calculates p-value by testing independence with the longest increasing
4 % subsequence, published by Jesus E. García and Verónica A. González-López
5 %
6 % Inputs: x - nx1 vector of data
7 %         y - nx1 vector of data
8 %         plot - if plot=1 you will get the distribution of D
9 % Output:
10 %         p - significance level
11 %         seq - longest increasing subsequence
12 %
13 % Marcus Vollmer
14 % 20.07.2010 - 22.07.2010
15
16 n=size(x,1);
17 [o TIEDADJ]=tiedrank(sortrows([x y]));
18 if TIEDADJ(1)~=0 %jitter x points if ranks are tied
19     x=x+normrnd(0,(max(x)-min(x))/n/1000,n,1);
20     o=tiedrank(sortrows([x y]));
21 end
22
23 %Calculation of the longest increasing subsequence (LIS)
24 [l,s]=LIS(o(:,2),0);
25 sortx=sort(x); sorty=sort(y);
26 seq=[sortx(s(:,1)) sorty(ceil(s(:,2)))];
27
28 %Calculation of two-sided significance level
29 if n>100 %using asymptotic distribution of L
30     T=(1-2*sqrt(n))/nthroot(n,6);
31     load('Matlab_SaveFiles/cdfTW2.mat');
32     if T<-3.9
33         p=cdfTW2(1);
34     elseif T>2.5
35         p=cdfTW2(size(cdfTW2,2));
36     else
37         p=2*min(cdfTW2(ceil((T+3.9)/0.01)),1-cdfTW2(floor((T+3.9)/0.01)));
38     end
39 else %using exact distribution of L
40     load('Matlab_SaveFiles/lisdf.mat');
41     a=2*sum(lisdf(n,1:l));
42     b=2*sum(lisdf(n,l:n));
43     p=min([1 a b]);

```

2 Von der Historie bis zur Gegenwart

```
44 end
45
46 %Plot
47 if plot==1
48     figure
49     scatter(x,y,'filled')
50     set(gcf,'PaperSize',[22 11],'position',[100 100 800 400],'color',[1 1
        1]);
51     set(gca,'xlim',[1.15*min(x)-0.15*max(x) 1.15*max(x)-0.15*min(x)],'ylim'
        ,[1.15*min(y)-0.15*max(y) 1.15*max(y)-0.15*min(y)],'box','on')
52     title('\fontsize{14}dataset','FontWeight','bold');
53     xlabel('\fontsize{14}x');
54     ylabel('\fontsize{14}y');
55     hold on
56     line(seq(:,1),seq(:,2),'LineStyle','-', 'Color','black','LineWidth',2)
57     scatter(seq(:,1),seq(:,2),'filled','red')
58     text(1.15*min(x)-0.15*max(x), 1.15*max(y)-0.15*min(y),['\bf\fontsize
        {10} p = ',sprintf('%0.3f',p), '\bf\fontsize{10} length = ',sprintf('
        %i',l)], 'HorizontalAlignment','left', 'VerticalAlignment','top', '
        BackgroundColor',[.85 .85 .85], 'Margin',3);
59     hold off
60 end
61 end
```

```
1 function [kmax seq]=LIS(z,output)
2 %Syntax: [kmax seq]=LIS(z,output)
3 % Calculates the longest increasing subsequence (LIS) of an sequence z
4 %
5 % Input:      z - nx1 vector of data
6 %            output - if true LIS will be displayed
7 % Output:    kmax - length of the LIS
8 %            seq - kmaxx2 vector (coordinates) of the LIS
9 %
10 % Marcus Vollmer
11 % 20.07.2010
12
13 n=size(z,1);
14 k=ones(n,1);
15
16 for i=2:n
17     for j=1:i-1
18         if z(j)<z(i)
19             if k(j)>=k(i)
20                 k(i)=k(j)+1;
21             end
22         end
23     end
24 end
```

```

25 kmax=max(k);
26
27 seq=zeros(kmax,2);
28 m=n;
29 temp=inf;
30 for i=0:kmax-1
31     for j=1:m
32         if k(j)==kmax-i && z(j)<temp
33             temp=z(j);
34             seq(kmax-i,1:2)=[j temp];
35             m=j-1;
36             break;
37         end
38     end
39 end
40
41 if output==1
42     writeData(z,n,kmax,seq);
43 end
44 end
45
46 %Output
47 function writeData(z,n,kmax,seq)
48 tr= repmat(' ',1,60);
49 disp('Input data: ')
50 disp(tr)
51 fprintf('Length: %i\nElements: ',n)
52 fprintf(' %i ',z)
53 fprintf('\n\nLongest Increasing Subsequence:\n')
54 disp(tr)
55 fprintf('Length: %i\nElements: ',kmax)
56 fprintf(' %i ',seq(:,2))
57 end

```

2.3.6 Geometrie zufälliger Permutationen

Der letzte Test, den wir nun betrachten wollen, ist bisweilen noch nicht veröffentlicht worden. Er entstammt einer Idee von CHRISTOPH BANDT und taucht erstmalig schriftlich in der Bachelorarbeit von KHALED M. AL ZOUKRA [48] auf. Der Test ist parameterfrei und untersucht die geometrische Anordnung von Datenpunkten. Nach dem englischen „Geometry of **R**andom **P**ermutations“ würden wir den Test auch mit „GRaP Independence Test“ bezeichnen. Zur Motivation betrachten wir die Ränge benachbarter Datenpunkte. Wir erkennen, dass diese nicht weit voneinander entfernt liegen, wenn die Merkmale abhängig sind. Bilden wir Umgebungen um die rangtransformierten Punkte, so erwarten wir bei Abhängigkeit viele, bei Unabhängigkeit weniger Überschneidungen, siehe Abbildung 2.15. Haben wir eine Beobachtung (X, Y) stetiger Variablen X und Y

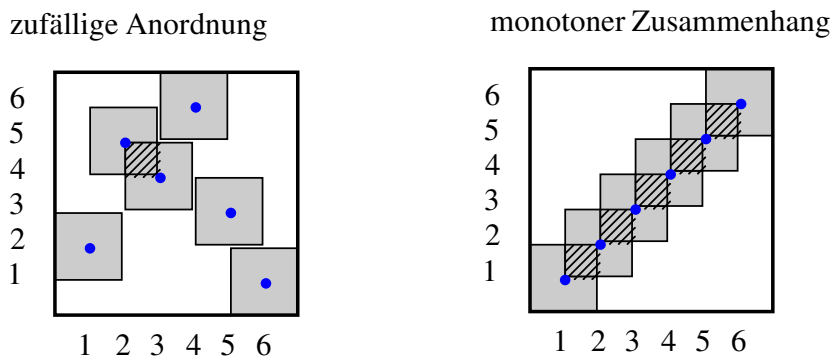


Abbildung 2.15: Links: Bei zufälliger Anordnung überschneiden sich nur wenige Flächen, Rechts: Aufgrund der Dichte an Punkten bei monotonen Zusammenhängen gibt es mehr Überschneidungen.

der reellen Zahlen, so entspricht deren Rangtransformierte einer Permutation π_s der Permutationsgruppe S_n . Unsere Beobachtung sei äquivalent zur Permutation π_s mit $\pi_s(i)=j_i$, $i=1, \dots, n$ und $j_i \in \{1, \dots, n\}$ mit $j_k \neq j_l$ für $k \neq l$. Wir nutzen die Matrixschreibweise einer Permutation:

$$P = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{pmatrix} \quad \text{mit} \quad p_{ij} = \begin{cases} 1, & \text{falls } \pi(i)=j, \\ 0, & \text{sonst.} \end{cases} \quad (2.47)$$

P wird als Permutationsmatrix bezeichnet. Legen wir um die rangtransformierten Punkte Quadrate mit Radius k , $k \in \mathbb{N}$, so lässt sich das Bild der Punkte, die von Quadraten überdeckt werden, ebenfalls als eine Matrix darstellen. $U(k)$ bezeichnen wir als

Punktüberdeckungsmatrix mit Radius k (halbe Seitenlänge der Quadrate). Sie sei wie folgt definiert:

$$U(k) = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{pmatrix} \quad \text{mit} \quad u_{ij} = \begin{cases} 1, & \text{falls } \exists l_1, l_2 \in \{-k, \dots, k\} \\ & \text{mit } \pi(i+l_1) = n-j+l_2, \\ 0, & \text{sonst.} \end{cases} \quad (2.48)$$

Einträgen mit dem Wert 1 entsprechen Punkte, die von mindestens einem Quadrat überdeckt werden. Einträgen mit Wert 0 entsprechen Punkte, die von keinem Quadrat überdeckt werden. Wie in Abbildung 2.16 zu erkennen, werden fast alle Punkte

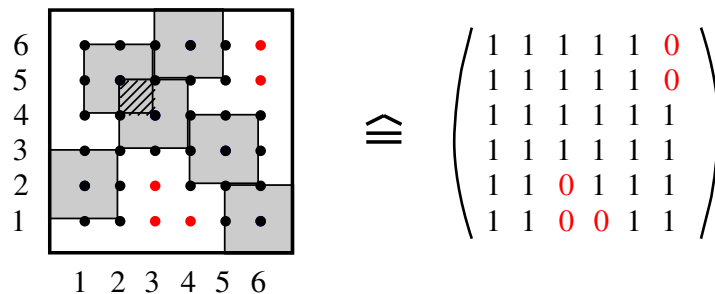


Abbildung 2.16: Links: schwarze Punkte werden von Quadraten überdeckt, rote Punkte sind überdeckungsfrei. Rechts: Überdeckungsmatrix - schwarze Punkte werden mit 1 kodiert, rote mit 0.

überdeckt. Wir erhalten eine Matrix, die viele Einsen enthält, besonders bei kleinen Stichproben. Nach diesem Prinzip werden einige freie Flächen, wie z.B. zwischen den Punkten (1, 3) und (2, 4) oder zwischen (4, 4) und (5, 5), nicht erfasst. Wir definieren daher die Flächenüberdeckungsmatrix F , mit $n'=n-1$:

$$F(k) = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1n'} \\ f_{21} & f_{22} & \dots & f_{2n'} \\ \vdots & \vdots & & \vdots \\ f_{n'1} & f_{n'2} & \dots & f_{n'n'} \end{pmatrix} \quad \text{mit} \quad f_{ij} = \begin{cases} 1, & \text{falls } \exists l_1, l_2 \in \{1-k, \dots, k\} \\ & \text{mit } \pi(i+l_1) = n-j+l_2, \\ 0, & \text{sonst.} \end{cases} \quad (2.49)$$

Abbildung 2.17 zeigt uns die Flächenüberdeckungsmatrix für das obige Beispiel. Wir erkennen, dass die Dimension der Matrix nun $(n-1) \times (n-1)$ ist. Dies entspricht nur den inneren Flächen, bzw. Quadraten, die zwischen den Punkten (1, 1) und (6, 6) liegen. Flächen außerhalb dieses Bereiches werden wir nicht betrachten, denn Randpunkte oder Eckpunkte haben weniger Nachbarn als Punkte des Inneren und werden deshalb weniger überdeckt

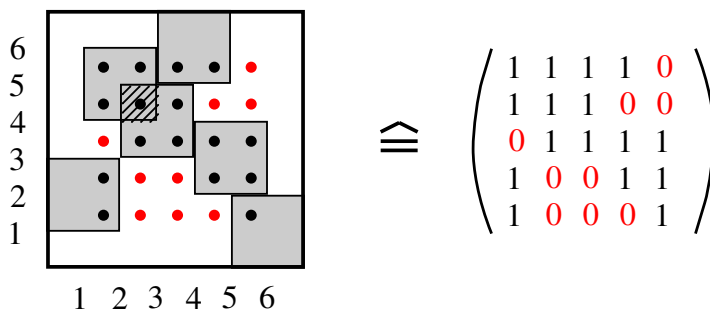


Abbildung 2.17: Links: schwarze Punkte werden von Quadraten überdeckt, rote Punkte stellen überdeckungsfreie Flächen dar. Rechts: Flächenüberdeckungsmatrix - schwarze Punkte werden mit 1 kodiert, rote mit 0.

als andere. Darauf werden wir später noch einmal eingehen.

Mit den Matrizen U und F eröffnen sich uns vier Möglichkeiten, die Abhängigkeit einer zufälligen Permutation zu messen. Zum Ersten durch die Wahl der Matrix und zum Zweiten entweder durch die Anzahl an Einsen oder durch die Anzahl an Nullen, die unsere Matrix enthält. Denn je weniger Einsen, umso mehr Überdeckungen haben wir. Je mehr Null-Einträge, umso mehr Überschneidungen muss es geben.

O.B.d.A. wählen wir die Möglichkeit über die Anzahl der bedeckten Punkte, also über die Anzahl der Einsen. Die Matrizen U und F werden wir simultan betrachten. Die Anzahl an Einsen der Überdeckungsmatrizen werden wir im Folgenden mit S_u und S_f bezeichnen. Sie sind definiert durch:

$$S_u = \sum_{(i,j) \in I_u} u_{ij} \quad S_f = \sum_{(i,j) \in I_f} f_{ij} \quad \text{mit} \quad (2.50)$$

$$I_u = [1+k, n-k] \times [1+k, n-k] \quad I_f = [k, n'-k+1] \times [k, n'-k+1]$$

I_u und I_f bezeichnen dabei den *inneren Bereich* der Punktüberdeckungs- bzw. Flächenüberdeckungsmatrix. In diesem Intervall hat jeder Punkt die gleiche Wahrscheinlichkeit überdeckt zu werden. Für jedes (i, j) gibt es gleich viele Paare (l_1, l_2) , die eine Überdeckung auslösen können. Abbildung 2.18 zeigt diese Anzahl. Punkte oder Flächen außerhalb dieses Intervalls werden dem *äußeren Bereich* zugeordnet.

Die Verteilung von S_u und S_f

Betrachten wir zufällige Permutationen der n -elementigen Menge $\{1, 2, \dots, n\}$, so werden wir leicht erkennen, dass die Anzahlen S_u und S_f einer Verteilung ähnlich der Binomial-

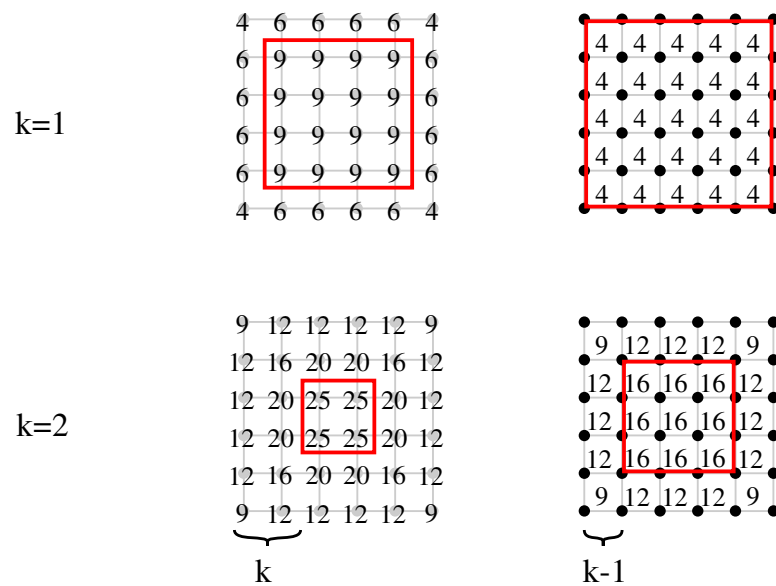


Abbildung 2.18: Beispiel des inneren Bereiches für $n=6$. In Zahlen jeweils die Anzahl an Punkten, welche eine Überdeckung auslösen können. Links: Punktüberdeckung, Rechts: Flächenüberdeckung. Die Punkte bzw. Flächen im rot eingerahmten inneren Bereich haben gleiche Anzahlen.

verteilung folgen. Im asymptotischen Fall erkennen wir das klare Bild einer Normalverteilung, vgl. dazu Abbildung 2.19. Doch es scheint nicht leicht zu sein, die exakte Verteilung anzugeben. Für kleinere n von 1 bis 9 können wir alle möglichen Permutationen generieren und die Summen S_u und S_f berechnen, um damit die exakte Verteilung angeben zu können. Für $n=9$ entspricht dies der Generierung von $9!=362880$ Permutationen. Die Berechnung der exakten Verteilung scheint für größere Umfänge aufgrund der mehr als exponentiell ansteigenden Laufzeit nicht möglich. Im Folgenden begnügen wir uns zuerst mit der Berechnung der Mittelwerte, den Minima und den Maxima beider Anzahlen.

Die Mittelwerte

Erklären wir uns zunächst, wie groß die Wahrscheinlichkeit ist, dass ein Element des inneren Bereiches mindestens einmal überdeckt wird. Dies ist nicht schwer zu ermitteln. Wir formulieren für die Überdeckung von Punkten: Sei (i, j) ein beliebiger Punkt des inneren Bereiches. Gesucht sei $P(u_{ij}=1)$. Über die Gegenwahrscheinlichkeit stellen wir auf: $P(u_{ij}=1) = 1 - P(u_{ij}=0)$. D.h. wir überlegen uns, wie groß die Wahrscheinlichkeit ist, dass (i, j) nicht überdeckt wird. Das heißt wiederum, dass kein Punkt in der Umgebung von (i, j) (mit Radius k) Teil der zufälligen Permutation ist. Also in jeder Spalte s mit $s=i-k, \dots, i+k$ ist $\pi(s) \notin \{j-k, \dots, j+k\}$. Die Wahrscheinlichkeit dafür ist:

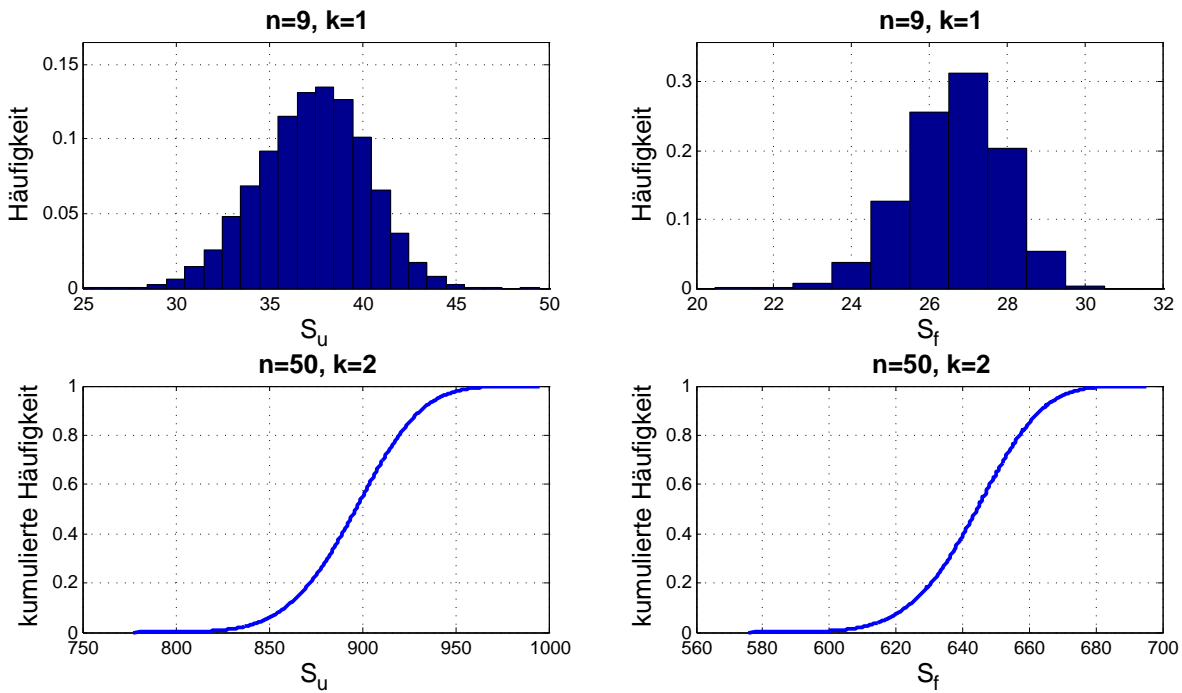


Abbildung 2.19: Oben: Exakte Dichteverteilung von S_u und S_f für $n=9$ und $k=1$. Unten: Verteilungsfunktion aus der Simulation für $n=50$ und $k=2$

$$\begin{aligned}
 P(u_{ij}=0) &= \frac{n-2k-1}{n} \cdot \frac{n-2k-2}{n-1} \cdot \dots \cdot \frac{n-2k-(2k+1)}{n-2k} \\
 &= \frac{(n-2k-1)^{2k+1}}{n^{2k+1}} \tag{2.51}
 \end{aligned}$$

Also einem Quotienten von fallenden Faktoriellen. Analog ergibt sich die Wahrscheinlichkeit, dass eine beliebige Fläche des inneren Bereiches nicht überdeckt wird:

$$\begin{aligned}
 P(f_{ij}=0) &= \frac{n-2k}{n} \cdot \frac{n-2k-1}{n-1} \cdot \dots \cdot \frac{n-2k-(2k-1)}{n-(2k-1)} \\
 &= \frac{(n-2k)^{2k}}{n^{2k}} \tag{2.52}
 \end{aligned}$$

Die Mittelwerte ergeben sich dann aus dem Produkt über der Anzahl der Elemente des inneren Bereiches mit der Wahrscheinlichkeit, dass ein Punkt (bzw. Feld) mindestens

einmal überdeckt wird:

$$\begin{aligned} M(S_u) &= (n-2k)^2 \cdot (1-P(u_{ij}=0)) = (n-2k)^2 \cdot \left(1 - \frac{(n-2k-1)^{2k+1}}{n^{2k+1}}\right) \\ &= (n-2k)^2 - \frac{(n-2k)^{2k+2}}{n^{2k}} \end{aligned} \quad (2.53)$$

$$\begin{aligned} M(S_f) &= (n-2k+1)^2 \cdot (1-P(f_{ij}=0)) = (n-2k+1)^2 \cdot \left(1 - \frac{(n-2k)^{2k}}{n^{2k}}\right) \\ &= (n-2k+1)^2 - \frac{(n-2k+1)^{2k+1}}{n^{2k-1}} \end{aligned} \quad (2.54)$$

Die Minima

Als nächstes wollen wir das Minimum herleiten. Dazu überlegen wir uns, wie die Punkte angeordnet sein müssen, damit möglichst wenig Fläche im inneren Bereich des Koordinatensystems entsteht. Möglichst wenig Fläche heißt, dass sich möglichst viele Umgebungen mit Radius k überschneiden. Spontan würden wir annehmen, dass eine lineare Anordnung von Punkten der beste Weg sei, um die minimale Anzahl von S_u und S_f zu bestimmen. Wir prüfen unsere Vermutung für $n \leq 9$. Wir erzeugen alle möglichen Permutationen und betrachten die Permutationen, die das Minimum von S_u und S_f bilden. Wir stellen fest, dass unsere „spontane“ Vermutung falsch war. Es zeigt sich, dass das Minimum von anderen Permutationen erzeugt wird. Sie sind der Art $\pi(1, \dots, n) = (\frac{n}{2}, \frac{n}{2}+1, \dots, n, 1, 2, \dots, \frac{n}{2}-1)$ für gerade n und $\pi(1, \dots, n) = (\lceil \frac{n}{2} \rceil, \lceil \frac{n}{2} \rceil + 1, \dots, n, 1, 2, \dots, \lfloor \frac{n}{2} \rfloor)$ für ungerade n , oder topologisch äquivalente Permutationen. Abbildung 2.20 zeigt sehr deutlich, dass wir durch Berechnung der L-förmigen Flächen schnell eine allgemeine Berechnungsvorschrift für das gesuchte Minimum angeben können. So gilt für $n \geq 8k$:

$$\min(S_u) = \left(\left\lceil \frac{n}{2} \right\rceil - 2k\right) (4k+1) + \left(\left\lfloor \frac{n}{2} \right\rfloor - 2k\right) (4k+1) = (n-4k)(4k+1) \quad (2.55)$$

In ähnlicher Vorgehensweise erhalten wir das Minimum von S_f für $n \geq 8k-2$:

$$\min(S_f) = \left(\left\lceil \frac{n}{2} \right\rceil - 2k+1\right) (4k-1) + \left(\left\lfloor \frac{n}{2} \right\rfloor - 2k+1\right) (4k-1) = (n-4k+2)(4k-1) \quad (2.56)$$

Erst durch die Bedingungen $n \geq 8k$ für S_u und $n \geq 8k-2$ für S_f stellen wir sicher, dass die beiden eingekreisten Bereiche der Abbildung 2.20 übereinstimmen. Zum Vergleich: Würden wir eine lineare Anordnung der Punkte vornehmen, so würden $n+4k(n-1-3k)$

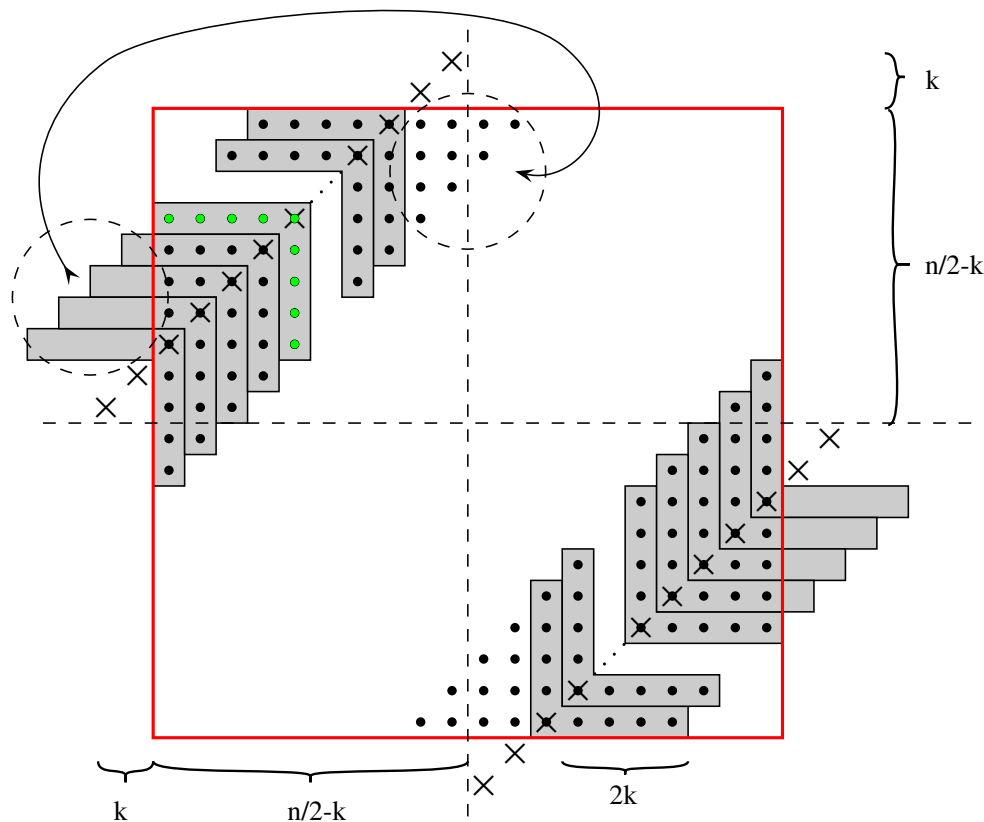


Abbildung 2.20: Koordinaten der Permutation π mit minimaler Anzahl S_u als Kreuze gekennzeichnet. Der innere Bereich ist rot umrandet. Schwarze Punkte stellen die überdeckten Koordinaten dar. Durch Auszählung der markierten Flächen ergibt sich das Minimum.

Punkte und $4k(n+2-3k)-n-1$ Flächen überdeckt werden. Das entspricht $4k^2$ mehr Punkten und $4k(k-1)+1$ mehr Flächen als beim Minimum.

Die Maxima

Haben wir keine Überschneidungen und liegen möglichst wenig Flächen im äußeren Bereich, so erhalten wir die Maxima. Da alle Rangzahlen von 1 bis n vorkommen müssen und es daher unumgänglich ist, dass Flächen im äußeren Bereich liegen, sollten wir mit dem Punkt $(1, 1)$ beginnen. So werden gleich 2 ungünstige Rangzahlen belegt. Es kann nun ein Punkt mehr im inneren Bereich liegen. Genauso ist es sinnvoll mit (n, n) zu enden. Von diesen Punkten aus konstruieren wir schrittweise eine Anordnung, die zum Maximum führt. Für $\max(S_u)$ bedeutet dies, dass der nächste Punkt $2k + 1$ Ränge entfernt liegen sollte,

damit keine Mehrfachüberdeckungen auftreten. Denn Mehrfachüberdeckungen bewirken eine Minderung der Summen S_u und S_f . Der nächste Punkt sollte demnach $(2, 2k + 2)$ sein - bei Flächenüberdeckung $(2, 2k + 1)$. Die Anordnung wird so vervollständigt, dass sich eine Anordnung wie in Abbildung 2.21 ergibt. Es entstehen $2k + 1$ aufsteigende Teil-

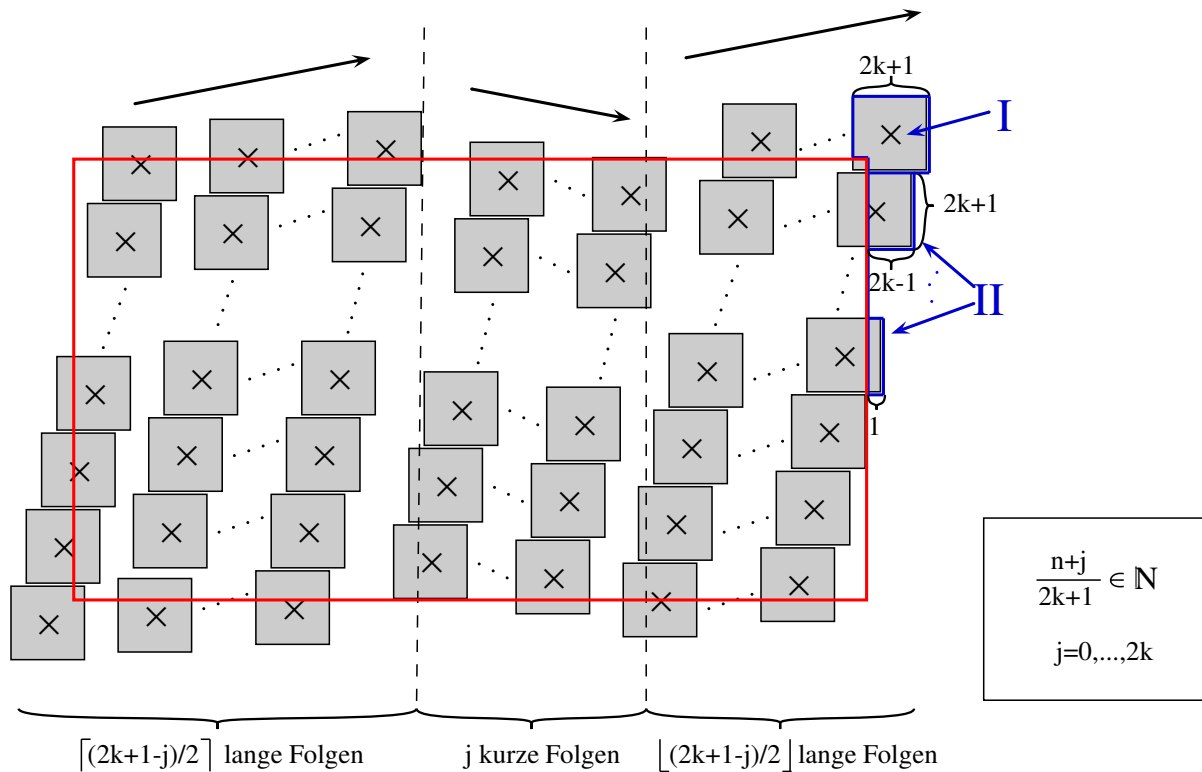


Abbildung 2.21: Schematische Darstellung der Anordnung von Punkten, deren Überdeckungen im inneren Bereich (rot) zum Maximum führt. Überdeckungen des äußeren Bereiches sind nicht relevant. Mehrfachüberdeckungen im inneren Bereich sollten nur ein mal gezählt werden - kommen aber aufgrund der besonderen Anordnung nur zu stand, wenn alle möglichen Punkte des inneren Bereiches überdeckt werden.

folgen, deren Längen zusammen n ergeben. Wenn n durch $2k+1$ teilbar ist, so enthält jede Folge gleich viele Punkte. Falls nicht und ist $n+j$ teilbar durch $2k+1$, so enthalten j Folgen jeweils einen Punkt weniger als die anderen $2k+1-j$ Folgen. Wir unterscheiden daher in kurze Folgen und lange Folgen (siehe Abb. 2.21). Sie sind so anzuordnen, dass zuerst $\lceil \frac{2k+1-j}{2} \rceil$ lange Folgen aufsteigend anzuordnen sind. Dann folgen die j kurzen Folgen in absteigender Weise. Der dritte Teil besteht dann aus $\lfloor \frac{2k+1-j}{2} \rfloor$ langen Folgen, welche wieder aufsteigend anzuordnen sind. Durch Wahl dieser Anordnung liegen am wenigsten

2 Von der Historie bis zur Gegenwart

Flächen im äußeren Bereich und es kann nur zu Mehrfachüberschneidungen zwischen den Folgen kommen - nicht innerhalb einer Folge. Doch wenn sich zwei Folgen überschneiden sollten, so folgt, dass alle Punkte des inneren Bereiches überdeckt sind. Dazu machen wir uns klar: Wenn sich 2 Folgen überschneiden, so gibt es keine Lücke, d.h. es existiert kein Punkt zwischen den Folgen, der nicht überdeckt ist. Wir stellen uns die Frage, in welchen Fällen sich zwei benachbarte Folgen schneiden. Dieser Fall tritt nur dann auf, wenn die Folgen aus zu wenig Elementen bestehen, genau genommen falls die Länge der langen Folgen $L_l = \frac{n+j}{2k+1}$ kleiner ist als $2k+2$. Die Folgen liegen dann dichter beieinander. Besonders kürzere Folgen liegen in x -Richtung einen Punkt dichter beieinander, als die langen Folgen. Durch die absteigende Anordnung der kürzeren Folgen tritt eine Überschneidung erst auf, wenn die Länge der kurzen Folgen $L_k = \frac{n+j}{2k+1} - 1$ kleiner ist als $2k+1$. Das ist die selbe Ungleichung wie für lange Folgen. Für $L_l < 2k+2$ sind dann auch alle Punkte des inneren Bereiches links der ersten Folge und rechts der zweiten Folge überdeckt. Falls es demnach Überschneidungen geben sollte, so überschneiden sich alle Folgen und der komplette innere Bereich wird überdeckt. Durch die beschriebene Anordnung sind also entweder alle Punkte des inneren Bereiches überdeckt oder aber es gibt keine Mehrfachüberschneidungen. Aufgrund der Minimalität der Flächen im äußeren Bereich folgt dann das gesuchte Maximum an Punktüberdeckungen.

Wir fassen alle Überlegungen zusammen und notieren den allgemeinen Fall. Sei

$$\frac{n+j}{2k+1} \in \mathbb{N} \quad \text{und} \quad j=0, \dots, 2k-1,$$

so überdecken die Flächen I,II der Abbildung 2.21 folgende Anzahl an Punkten:

$$I = (2k+1)^2 - 1 = 4k^2 + 4k$$

$$II = (2k+1) \cdot ((2k-1) + \dots + 2+1) = (2k+1)(2k) \left(\frac{2k-1}{2} \right) = k(4k^2 - 1)$$

Für den Fall, dass $\frac{n+j}{2k+1} < 2k+2$, ist der komplette innere Bereich überdeckt und es folgt :

$$\max(S_u) = (n-2k)^2$$

Andernfalls gibt es keine Mehrfachüberdeckungen und wir überlegen, dass jeder der n Punkte maximal $(2k+1)^2$ Punkte überdecken kann. Dazu zählen aber auch Flächen des äußeren Bereiches. Diese werden von den k höchsten und k niedrigsten Rängen erzeugt. Die blau markierten Flächen der Abbildung 2.21 mit Markierung I sind daher 2 mal und die Flächen mit Markierung II sind 4 mal abzuziehen.

$$\begin{aligned}
 \max(S_u) &= n \cdot (2k+1)^2 - 2 \cdot I - 4 \cdot II \\
 &= n(2k+1)^2 - 2(4k^2+4k) - 4k(4k^2-1) \\
 &= n(2k+1)^2 - 4k(1+2k+4k^2)
 \end{aligned}$$

Einen besonderen Fall haben wir für $j=2k$. Wir müssen dann eine lange Folge und $2k$ kurze Folgen anordnen. Beginnen wir mit $(1, 1)$, so würde die erste Folge (lange Folge) bei $(\frac{n+j}{2k+1}, n)$ enden. Das würde bedeuten, dass der Punkt (n, n) nicht möglich wäre. Dieser ist aber wichtig, damit möglichst wenig überdeckte Punkte im äußeren Bereich liegen. Daher sollte die erste Folge eine kurze Folge sein und die letzte Folge sollte um den Punkt (n, n) erweitert werden. Für $n=31$ und $k=2$ ergibt sich folgendes Bild. Wie wir erkennen können,

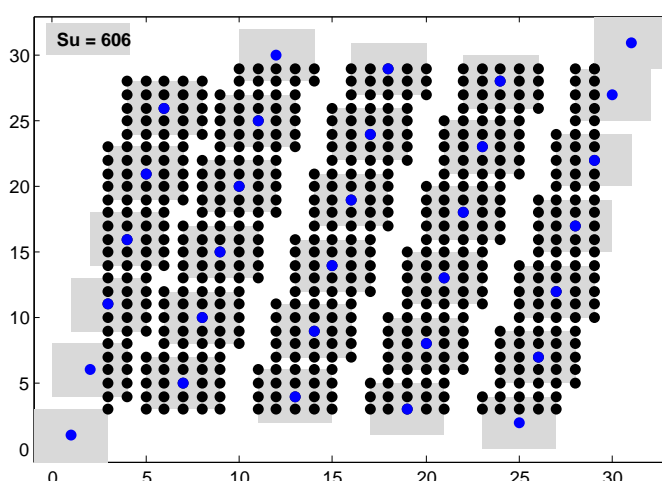


Abbildung 2.22: Anordnung für den Fall $j=2k$. Hier das Beispiel für $n=31$ und $k=2$. Die letzte kurze Folge wird um den Punkt (n, n) erweitert. Dabei gelangen $(2k+1) \cdot (2k)$ Punkte der letzten Folge in den inneren Bereich. Dem gegenüber steht nur ein Verlust von $(2k+1)$ Punkten, die ursprünglich den Abschluss der ersten Folgen bildeten.

liegen die ersten beiden Folgen nun dichter beieinander als die restlichen Folgen. Daher sind für geringere n Mehrfachüberschneidungen noch vor der kompletten Abdeckung des inneren Bereiches möglich. Dann führt wiederum erst genanntes Anordnungsprinzip zum Maximum. Mehrfachüberschneidungen ohne Komplettüberdeckung des inneren Bereiches gibt es nur im Falle $\frac{n+2k}{2k+1} = 2k+2$. Komplettüberdeckungen für $\frac{n+2k}{2k+1} < 2k+2$. So können wir für $\frac{n+2k}{2k+1} > 2k+2$ und $j=2k$ notieren:

$$\begin{aligned}
 \max(S_u) &= n(2k+1)^2 - 2(2k+1)^2 + 1 - 4k(4k^2-1) \\
 &= (n-2)(2k+1)^2 + 1 - 4k(4k^2-1)
 \end{aligned}$$

2 Von der Historie bis zur Gegenwart

Zusammen ergibt dies also

$$\max(S_u) = \begin{cases} (n-2)(2k+1)^2 + 1 - 4k(4k^2-1) & , \text{für } \frac{n+j}{2k+1} > 2k+2 \text{ und } j=2k \\ (n-2k)^2 & , \text{für } \frac{n+j}{2k+1} < 2k+2 \\ n(2k+1)^2 - 4k(1+2k+4k^2) & , \text{sonst} \end{cases} \quad (2.57)$$

$$\text{mit } \frac{n+j}{2k+1} \in \mathbb{N} \quad \text{und} \quad j=0, \dots, 2k$$

In analoger Weise erhalten wir das Maximum für die Flächenüberdeckung, mit $\frac{n+j}{2k} \in \mathbb{N}$ und $j=0, \dots, 2k-1$:

$$\max(S_f) = \begin{cases} 1 - 4k(4k^2 - nk - 4k + 2) & , \text{für } \frac{n+j}{2k} > 2k+1 \text{ und } j=2k-1 \\ (n-2k+1)^2 & , \text{für } \frac{n+j}{2k} < 2k+1 \\ 2 - 4k(4k^2 - nk - 4k + 2) & , \text{sonst} \end{cases} \quad (2.58)$$

Asymptotische Verteilungen

Wie schon bemerkt, fällt es uns schwer, die exakte Verteilung von S_u und S_f anzugeben. Doch wie wir bereits aus Abbildung 2.19 schließen konnten, lassen sich unsere Verteilungen sehr gut mittels Normalverteilungen anpassen. Abbildung 2.23 zeigt uns, wie gut die Anpassung von S_u durch eine Normalverteilung wirklich ist. Die Verteilung von S_u in rot

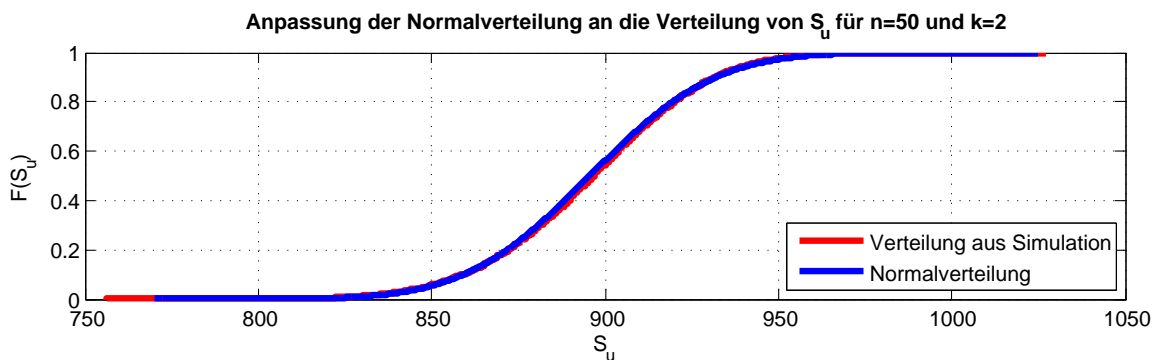


Abbildung 2.23: Anpassung einer Normalverteilung an die Verteilung von S_u . Hier das Beispiel für $n=50$ und $k=2$.

ist dabei aus einer Simulation von 500000 zufälligen Permutationen für $n=50$ und $k=2$ entstanden. Die Normalverteilung in blau zeigt nur geringe Unterschiede an den Enden

der Verteilung. Für die verschiedenen Stichprobenumfänge n und Radien k benötigen wir lediglich die Mittelwerte und die Streuungen der anzupassenden Verteilungen von S_u und S_f . Die Mittelwerte bestimmen wir mittels der Formeln (2.53) und (2.54). Lediglich die Standardabweichungen müssen aus Simulationen heraus geschätzt werden. Schauen wir uns die Streuungen für verschiedene Stichprobenumfänge an. Folgende Abbildung zeigt uns die Streuungen für $n = 10, \dots, 100$, welche aus jeweils 100000 zufälligen Permutationen hervorgegangen sind. Wir stellen fest, dass sich die Streuungen leicht durch ein

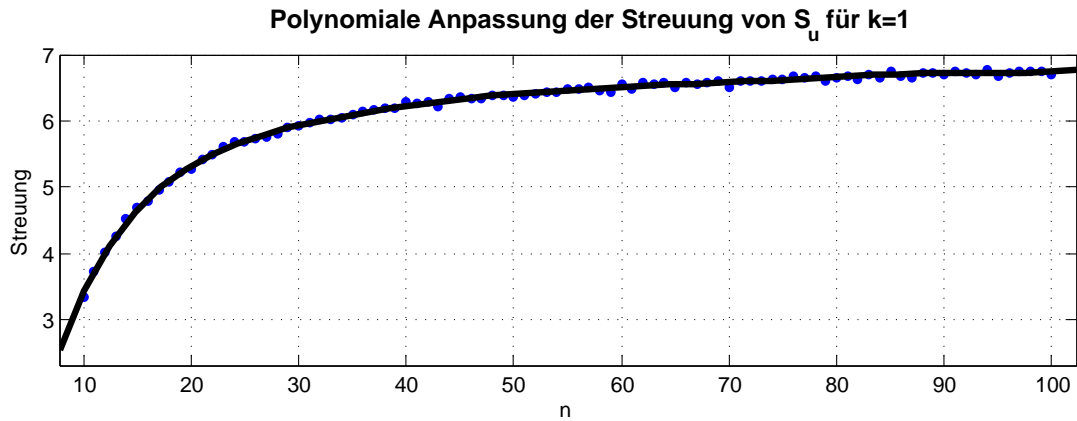


Abbildung 2.24: Anpassung der Streuung durch ein Polynom 7. Grades.

Polynom anpassen lassen. Folgende Polynome liefern für die Approximation der Streuung von S_u und S_f sehr gute Ergebnisse:

	Polynome zur Streuung von S_u für $n=10, \dots, 100$
$k=1$	$\hat{\sigma}(S_u) = 4,9387 \cdot 10^{-12} n^7 - 2,0955 \cdot 10^{-9} n^6 + 3,6894 \cdot 10^{-7} n^5 - 3,4905 \cdot 10^{-5} n^4 + 1,9214 \cdot 10^{-3} n^3 - 0,0623 \cdot n^2 + 1,1412 n - 3,4252$
$k=2$	$\hat{\sigma}(S_u) = -1,0348 \cdot 10^{-6} n^4 + 3,2089 \cdot 10^{-4} n^3 - 0,0382 \cdot n^2 + 2,1882 n - 19,1250$
$k=3$	$\hat{\sigma}(S_u) = 2,3676 \cdot 10^{-9} n^6 - 9,1404 \cdot 10^{-7} n^5 + 1,4212 \cdot 10^{-4} n^4 - 1,1252 \cdot 10^{-2} n^3 + 0,4601 \cdot n^2 - 7,4254 n + 39,5242$
$k=4$	$\hat{\sigma}(S_u) = -9,8622 \cdot 10^{-8} n^5 + 3,6282 \cdot 10^{-5} n^4 - 5,2140 \cdot 10^{-3} n^3 + 0,3528 \cdot n^2 - 8,9135 n + 73,5257$

Die Abweichung der polynomiell berechneten Streuungen zu den Streuungen der Simulation liegen in der Größenordnung von bis zu 1%. Die geringe Abweichung hat daher nur einen verschwindend geringen Einfluss auf den p-Wert.

2 Von der Historie bis zur Gegenwart

Polynome zur Streuung von S_f für $n=10, \dots, 100$	
$k=1$	$\hat{\sigma}(S_f) = 3,5814 \cdot 10^{-10} n^5 - 1,1393 \cdot 10^{-7} n^4 + 1,3988 \cdot 10^{-5} n^3 - 8,3014 \cdot 10^{-4} n^2 + 0,0242 n + 1,1128$
$k=2$	$\hat{\sigma}(S_f) = -3,6961 \cdot 10^{-10} n^6 + 1,4366 \cdot 10^{-7} n^5 - 2,2808 \cdot 10^{-5} n^4 + 1,9071 \cdot 10^{-3} n^3 - 0,0903 \cdot n^2 + 2,4162 n - 14,3014$
$k=3$	$\hat{\sigma}(S_f) = 1,8840 \cdot 10^{-9} n^6 - 6,8549 \cdot 10^{-7} n^5 + 9,8732 \cdot 10^{-5} n^4 - 7,0410 \cdot 10^{-3} n^3 + 0,2448 \cdot n^2 - 2,5568 n + 5,1841$
$k=4$	$\hat{\sigma}(S_f) = -1,2839 \cdot 10^{-7} n^5 + 4,2966 \cdot 10^{-5} n^4 - 5,5133 \cdot 10^{-3} n^3 + 0,3250 \cdot n^2 - 6,8246 n + 45,7477$

Die genauen Koeffizienten der Polynome, sowie die Residuen, also der Abweichung zu den Streuungen der Simulation, können den Matlab-Dateien `PolynomialeAnpassung100Sf.mat` und `PolynomialeAnpassung100Su.mat` entnommen werden. Die Dateien befinden sich auf der mitgelieferten Daten-CD im Ordner `\Matlab\Matlab_SaveFiles`. Für $100 \leq n \leq 1000$ können folgende Polynome als Näherung für die Streuung von S_u und S_f herangezogen werden:

Polynome zur Streuung von S_u für $n=100, \dots, 1000$	
$k=1$	$\hat{\sigma}(S_u) = 1.3194 \cdot 10^{-19} n^7 - 5.1770 \cdot 10^{-16} n^6 + 8.4379 \cdot 10^{-13} n^5 - 7.4033 \cdot 10^{-10} n^4 + 3.7748 \cdot 10^{-7} n^3 - 1.1199 \cdot 10^{-4} n^2 + 1.8163 \cdot 10^{-2} n + 5.7080$
$k=2$	$\hat{\sigma}(S_u) = 3.4835 \cdot 10^{-18} n^7 - 1.3956 \cdot 10^{-14} n^6 + 2.2835 \cdot 10^{-11} n^5 - 1.9663 \cdot 10^{-8} n^4 + 9.5794 \cdot 10^{-6} n^3 - 2.6463 \cdot 10^{-3} n^2 + 0.3946 n + 14.1736$
$k=3$	$\hat{\sigma}(S_u) = 1.0820 \cdot 10^{-17} n^7 - 4.4411 \cdot 10^{-14} n^6 + 7.4992 \cdot 10^{-11} n^5 - 6.7292 \cdot 10^{-8} n^4 + 3.4616 \cdot 10^{-5} n^3 - 1.0280 \cdot 10^{-2} n^2 + 1.6896 n - 10.8814$
$k=4$	$\hat{\sigma}(S_u) = 2.0206 \cdot 10^{-17} n^7 - 8.3551 \cdot 10^{-14} n^6 + 1.4280 \cdot 10^{-10} n^5 - 1.3067 \cdot 10^{-7} n^4 + 6.9402 \cdot 10^{-5} n^3 - 2.1734 \cdot 10^{-2} n^2 + 3.9159 n - 94.7707$

Polynome zur Streuung von S_f für $n=100, \dots, 1000$	
$k=1$	$\hat{\sigma}(S_f) = -1,5008 \cdot 10^{-13} n^4 + 4,2469 \cdot 10^{-10} n^3 - 4,1762 \cdot 10^{-7} n^2 + 1,6750 \cdot 10^{-4} n + 1,3906$
$k=2$	$\hat{\sigma}(S_f) = 1,0266 \cdot 10^{-18} n^7 - 4,1324 \cdot 10^{-15} n^6 + 6,7846 \cdot 10^{-12} n^5 - 5,8519 \cdot 10^{-9} n^4 + 2,8502 \cdot 10^{-6} n^3 - 7,8542 \cdot 10^{-4} n^2 + 0,1164 n + 11,6149$

	Polynome zur Streuung von S_f für $n=100, \dots, 1000$
$k=3$	$\hat{\sigma}(S_f) = 5,9403 \cdot 10^{-18} n^7 - 2,4301 \cdot 10^{-14} n^6 + 4,0837 \cdot 10^{-11} n^5 - 3,6396 \cdot 10^{-8} n^4$ $+ 1,8540 \cdot 10^{-5} n^3 - 5,4232 \cdot 10^{-3} n^2 + 0,8688 n + 9,6054$
$k=4$	$\hat{\sigma}(S_f) = 1,7196 \cdot 10^{-17} n^7 - 7,0253 \cdot 10^{-14} n^6 + 1,1830 \cdot 10^{-10} n^5 - 1,0618 \cdot 10^{-7} n^4$ $+ 5,4917 \cdot 10^{-5} n^3 - 1,6537 \cdot 10^{-2} n^2 + 2,7999 n - 48,4885$

Sie sind ebenfalls unter `\Matlab\Matlab_SaveFiles` abrufbar. Die Koeffizienten (gespeichert als `beta0` bis `beta12` für $k=1, \dots, 13$) befinden sich in `PolynomialeAnpassung1000Su.mat` und `PolynomialeAnpassung1000Sf.mat`.

Der p-Wert

Zur Entscheidung, ob ein Datensatz einen Zusammenhang in sich birgt, können wir den p-Wert für $n \leq 100$ entweder durch „Ablesen“ des Quantils aus einer Quantilstabelle ermitteln, oder mit Hilfe der Normalverteilung berechnen. Um die Quantile zu bestimmen, haben wir für $n \leq 9$ die exakte Verteilung berechnet. Für $10 \leq n \leq 100$ dienen zur Bestimmung der Verteilungen Simulationen von je 500000 zufälligen Permutationen. Die Quantile werden dabei sehr exakt geschätzt. Für $k=1$ weicht der Durchschnitt der berechneten Quantile für S_u mit 95%-iger Wahrscheinlichkeit maximal um 0.0625 ab. Für S_f sogar nur um maximal 0.0288. Die resultierenden Tabellen mit den wichtigsten Quantilen für $k=1, \dots, 4$ befinden sich im Anhang (5.7). Entscheiden wir uns für den Weg über die Normalverteilung, so berechnen wir den Mittelwert nach (2.53) oder (2.54) und schätzen die Varianz durch die angegebenen Polynome. Der p-Wert ergibt sich aus dem einseitigen Ablehnungsbereich für die Normalverteilung. Für $n > 100$ sollten wir zur Berechnung des p-Wertes grundsätzlich die Normalverteilung nutzen, da die Berechnung von Quantilen für diese Größenordnungen zu zeitintensiv ist und weil die Normalverteilung ein nahezu perfekte Approximation liefert.

Haben wir demnach eine Stichprobe (x, y) des Umfangs n und resultieren die empirischen Summen $s_u(x, y)$ und $s_f(x, y)$ aus der Punktüberdeckung bzw. Flächenüberdeckung, so berechnet sich der p-Wert aus der kumulierten Verteilungsfunktion Φ der Standard-Normalverteilung wie folgt:

$$\hat{p}_u = \Phi \left(\frac{s_u(x, y) - \mu_u}{\hat{\sigma}(S_u)} \right) \quad \hat{p}_f = \Phi \left(\frac{s_f(x, y) - \mu_f}{\hat{\sigma}(S_f)} \right) \quad (2.59)$$

Die Mittelwerte μ_u und μ_f berechnen sich dabei aus (2.53) und (2.54). Die Varianz wird durch obige Polynome geschätzt.

Die richtige Wahl des Radius k

Wir möchten uns als nächstes der richtigen Wahl des Radius k widmen. Es lassen sich sehr leicht Beispiele konstruieren, bei denen der Geometrietest einen Zusammenhang nicht erkennen kann, wenn die Umgebungsgröße k zu klein gewählt wird. Besonders bei größeren Stichprobenumfängen liegen die Punkte i.d.R. weiter als 2 Ränge auseinander - auch bei zusammenhängenden Merkmalen. Überdeckungen kommen in solchen Fällen nur dann zu Stande, wenn der Radius k vergrößert wird. Abbildung 2.25 zeigt ein solches Beispiel.

Bei einem Radius $k=1$ treten wenig Überdeckungen auf – zu wenig um eine Abhängigkeit

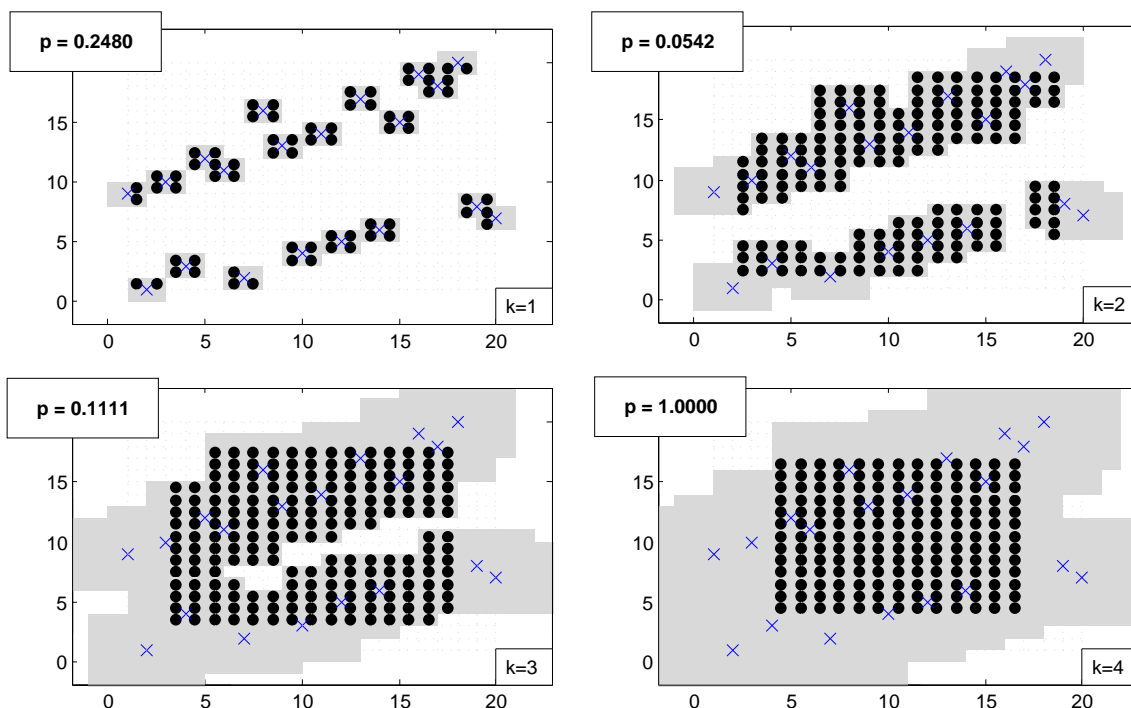


Abbildung 2.25: Verschiedene Signifikanzwerte bei unterschiedlicher Wahl des Radius k . Hier ein Beispiel für einen doppelten linearen Zusammenhang. Für $k=1$ kann noch kein signifikanter Unterschied zur zufälligen Anordnung von Datenpunkten ausgemacht werden. $k=2$ liefert den besten p-Wert.

zu detektieren. Doch schon bei Vergrößerung des Radius auf $k=2$ erhalten wir signifikant viele Überschneidungen. Die Abhängigkeit scheint bewiesen. Eine weitere Erhöhung des Radius zeigt jedoch einen Abfall des p-Wertes. Zu erklären ist dies mit der gleichzeitigen Zunahme der Überschneidungen bei zufälligen Permutationen. Überschneidungen werden zur Normalität. Die Entscheidung auf Abhängigkeit oder Unabhängigkeit zweier stetiger Merkmale hängt somit entscheidend von der Wahl des Radius k ab. Doch welcher Radius

ist für welchen Stichprobenumfang am Besten geeignet? Nach obigem Beispiel würden wir urteilen, dass für große Stichprobenumfänge ein kleines k nicht geeignet ist. Selbst abhängige Daten würden zu wenig Überschneidungen aufwerfen. Auch die Verteilungen von S_u und S_f scheinen dann nicht mehr normalverteilt zu sein. Durch den Drang zu wenigen Mehrfachüberdeckungen scheinen die Verteilungen dem Minimum näher als dem Maximum. Es schleicht sich eine Schiefe ein - die Verteilung ist linkssteil. Andererseits haben wir den gleichen ungünstigen Effekt, würden wir den Radius k zu groß wählen. Dann ist unsere Verteilung rechtssteil. Gerade für die approximative Berechnung des p-Wertes durch die Normalverteilung ist dies nicht akzeptabel. k muss also so gewählt werden, dass nicht zu wenig aber auch nicht zu viele Punkte (bzw. Flächen) des inneren Bereiches abgedeckt werden. Wir entscheiden uns dafür, dass im Durchschnitt die Hälfte aller inneren Punkte (Flächen) abgedeckt werden. Das heißt:

$$\frac{M(S_u)}{(n-2k)^2} = 1 - \frac{(n-2k-1)^{2k+1}}{n^{2k+1}} \approx \frac{1}{2} \quad (2.60)$$

$$\frac{M(S_f)}{(n-2k+1)^2} = 1 - \frac{(n-2k)^{2k}}{n^{2k}} \approx \frac{1}{2} \quad (2.61)$$

Für jeden Stichprobenumfang n lässt sich somit ein optimales k finden. Abbildung 2.26 zeigt uns jenes k , für welches im Durchschnitt die Hälfte aller inneren Punkte (Flächen) abgedeckt werden:

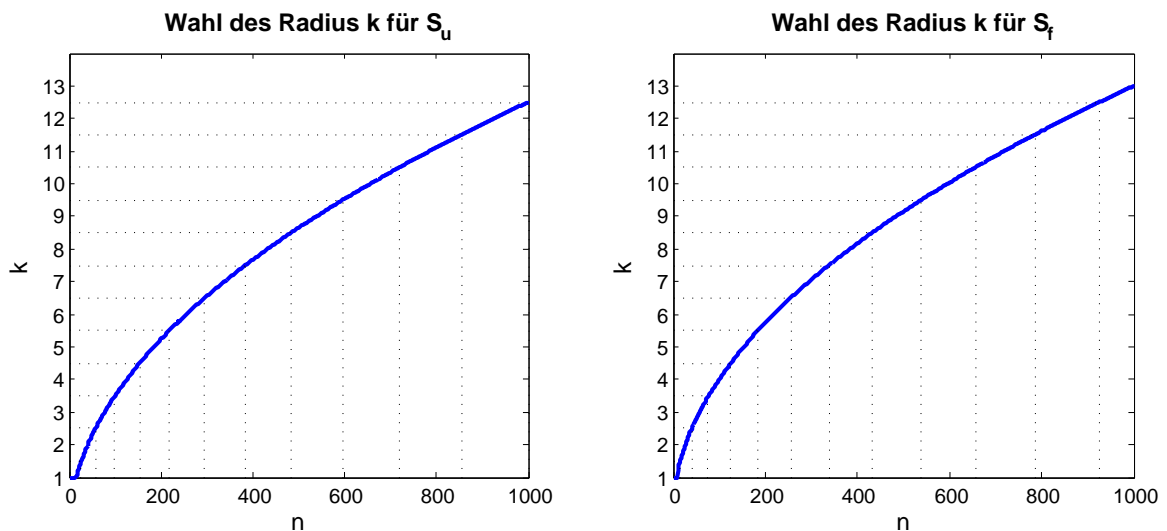


Abbildung 2.26: Approximativ bestimmtes k für $n=1, \dots, 1000$, als Lösung von (2.60) bzw. (2.61).

2 Von der Historie bis zur Gegenwart

Da wir k nur ganzzahlig wählen können, erhalten wir nach Rundung zur nächstgelegenen ganzen Zahl folgende Bereiche als Empfehlung zur Wahl des Radius:

S_u				S_f			
n	k	n	k	n	k	n	k
1...26	1	385...484	8	1...15	1	339...433	8
27...57	2	485...596	9	16...39	2	434...538	9
58...99	3	597...719	10	40...76	3	539...656	10
100...153	4	720...854	11	77...124	4	657...785	11
154...218	5	855...1000	12	125...184	5	786...925	12
219...295	6			185...255	6	926...1000	13
296...384	7			256...338	7		

Matlab [GRaP Independence Test]

Mit Eingabe der Stichprobenvektoren x , y erfolgt die Berechnung der Teststatistik S_u und S_f mit empfohlenem Radius k , sowie die Berechnung des Signifikanzniveaus unter Nutzung von Quantilstabellen oder für größere Stichprobenumfänge durch Nutzung der asymptotischen Normalverteilung. Optional kann k direkt vorgegeben werden. Ebenfalls möglich ist eine grafische Ausgabe.

```

1 function [Su Sf pu pf]=GRaPInd(x,y,varargin)
2 %Syntax: [Su Sf pu pf]=GRaPInd(x,y,varargin)
3 % Calculates p-value by testing independence by geometry of random
4 % permutations, invented by C. Bandt, K. M. al Zoukra and Marcus Vollmer
5 %
6 % Inputs: x - nx1 vector of data
7 %         y - nx1 vector of data
8 % optional Inputs:
9 %         k - radius of squares
10 %         plot - if plot=1 you will get the distribution of Su, Sf, and
11 %              pictures of area around points
12 % Output:
13 %         Su - sum of overlapping points
14 %         Sf - sum of overlapping areas
15 %         pu - significance level using overlapmatrix
16 %         pf - significance level using matrix of area
17 %
18 % Marcus Vollmer
19 % 28.07.2010 / 26.08.2010
20
21 switch nargin
22     case 2; load('Matlab_SaveFiles/GRaPOptimalK.mat');

```

```

23         k=sum(ksu<=size(x,1))+1; plot=0;
24     case 3; k=varargin{1}; plot=0;
25     case 4; k=varargin{1}; plot=varargin{2};
26     otherwise; fprintf('Input Failure! Try help GRaPInd')
27 end
28
29 n=size(x,1);
30 [o TIEDADJ]=tiedrank(sortrows([x y]));
31 if TIEDADJ(1)~=0 %jitter x points if ranks are tied
32     x=x+normrnd(0,(max(x)-min(x))/n/1000,n,1);
33     o=tiedrank(sortrows([x y]));
34 end
35 if TIEDADJ(2)~=0 %jitter y points if ranks are tied
36     y=y+normrnd(0,(max(y)-min(y))/n/1000,n,1);
37     o=tiedrank(sortrows([x y]));
38 end
39
40 %Calculation of Su (counting ones in the overlapmatrix)
41 % u - overlapmatrix, mu - multioverlapping
42 if plot==1
43     figure
44     if n<20
45         scatter(reshape(repmat(1+k:n-k,n-2*k,1),(n-2*k)^2,1),reshape(repmat(
46             (1+k:n-k,n-2*k,1),(n-2*k)^2,1),'filled','MarkerFaceColor',[1 0
47             0])
48         end
49         property1(n,k)
50     end
51     mu=zeros(n+2*k,n+2*k);
52     for i=1:n
53         x1=i-k; x2=i+k;
54         y1=o(i,2)-k; y2=o(i,2)+k;
55         mu(x1+k:x2+k,y1+k:y2+k)=mu(x1+k:x2+k,y1+k:y2+k)+1;
56         if plot==1
57             patch([x1 x1 x2 x2],[y1 y2 y2 y1],[1 1 1 1],'LineStyle','none','
58                 facecolor',[.85 .85 .85],'AlphaDataMapping','none')
59             x11=max(x1,k+1); x22=min(x2,n-k); dim1=x22-x11+1;
60             y11=max(y1,k+1); y22=min(y2,n-k); dim2=y22-y11+1;
61             scatter(reshape(repmat(x11:x22,dim2,1),dim1*dim2,1),reshape(repmat(
62                 y11:y22,dim1,1),dim1*dim2,1),'filled','MarkerFaceColor',[0 0
63                 0])
64         end
65     end
66     u=sign(mu(k+1:n+k,k+1:n+k));
67     Su=sum(sum(u(1+k:n-k,1+k:n-k)));
68     if plot==1
69         scatter(o(:,1),o(:,2),'filled','MarkerFaceColor',[0 0 1])
70         text(0,max(y)+1,['\bf\fontsize{10}\_Su=\_'],'sprintf('%i',Su),'
71             HorizontalAlignment','left','VerticalAlignment','top','
72             BackgroundColor',[.85 .85 .85],'Margin',3);
73         set(gca,'xlim',[1-k n+k],'ylim',[1-k n+k],'box','on')
74     hold off

```

2 Von der Historie bis zur Gegenwart

```

68 end
69
70 %Calculation of Sf (counting ones in the matrix of area)
71 % f - areamatrix, mf - multiarea
72 if plot==1
73     figure
74     if n<20
75         scatter(reshape(repmat(0.5+k:n+0.5-k,n-2*k+1,1),(n-2*k+1)^2,1),
76                 reshape(repmat(0.5+k:n+0.5-k,n-2*k+1,1)',(n-2*k+1)^2,1),'filled',
77                 'MarkerFaceColor',[1 0 0])
78     end
79     property1(n,k)
80 end
81 mf=zeros(n-1+2*k,n-1+2*k);
82 for i=1:n
83     x1=i-k; x2=i+k;
84     y1=o(i,2)-k; y2=o(i,2)+k;
85     mf(x1+k:x2+k-1,y1+k:y2+k-1)=mf(x1+k:x2+k-1,y1+k:y2+k-1)+1;
86     if plot==1
87         patch([x1 x1 x2 x2],[y1 y2 y2 y1],[1 1 1 1],'LineStyle','none','
88             facecolor',[.85 .85 .85])
89         x11=max(x1,k)+.5; x22=min(x2,n-k+1)-.5; dim1=x22-x11+1;
90         y11=max(y1,k)+.5; y22=min(y2,n-k+1)-.5; dim2=y22-y11+1;
91         scatter(reshape(repmat(x11:x22,dim2,1),dim1*dim2,1),reshape(repmat(
92             y11:y22,dim1,1)',dim1*dim2,1),'filled','MarkerFaceColor',[0 0
93             0])
94     end
95 end
96 f=sign(mf(k+1:n+k-1,k+1:n+k-1)); %outer interval included
97 Sf=sum(sum(f(k:n-k,k:n-k))); %only inner interval
98 if plot==1
99     scatter(o(:,1),o(:,2),'Marker','x','SizeData',10^2,'MarkerFaceColor',[0
100         0 1])
101     text(0,-1,['\bf\ fontsize{10}_Sf=_'],sprintf('%i',Sf),'
102         HorizontalAlignment','left','VerticalAlignment','top','
103         BackgroundColor',[.85 .85 .85],'Margin',3);
104 hold off
105 end
106
107 %Calculation of significance level
108 if n>100 %using asymptotic distribution of Su, Sf
109     muu=(n-2*k)^2-prod((n-4*k-1):(n-2*k))/prod((n-2*k+1):(n));
110     muf=(n-2*k+1)^2-prod((n-4*k+1):(n-2*k+1))/prod((n-2*k+2):(n));
111     polyu=load('Matlab_SaveFiles/PolynomialeAnpassung1000Su.mat',strcat('
112         beta',num2str(k-1)));
113     polyf=load('Matlab_SaveFiles/PolynomialeAnpassung1000Sf.mat',strcat('
114         beta',num2str(k-1)));
115     switch k
116     case 1; sdu=polyval(polyu.beta0,n); sdf=polyval(polyf.beta0,n);
117     case 2; sdu=polyval(polyu.beta1,n); sdf=polyval(polyf.beta1,n);
118     case 3; sdu=polyval(polyu.beta2,n); sdf=polyval(polyf.beta2,n);
119     case 4; sdu=polyval(polyu.beta3,n); sdf=polyval(polyf.beta3,n);

```



```

110     case 5; sdu=polyval(polyu.beta4,n); sdf=polyval(polyf.beta4,n);
111     case 6; sdu=polyval(polyu.beta5,n); sdf=polyval(polyf.beta5,n);
112     case 7; sdu=polyval(polyu.beta6,n); sdf=polyval(polyf.beta6,n);
113     case 8; sdu=polyval(polyu.beta7,n); sdf=polyval(polyf.beta7,n);
114     case 9; sdu=polyval(polyu.beta8,n); sdf=polyval(polyf.beta8,n);
115     case 10; sdu=polyval(polyu.beta9,n); sdf=polyval(polyf.beta9,n);
116     case 11; sdu=polyval(polyu.beta10,n); sdf=polyval(polyf.beta10,n);
117     case 12; sdu=polyval(polyu.beta11,n); sdf=polyval(polyf.beta11,n);
118     case 13; sdu=polyval(polyu.beta12,n); sdf=polyval(polyf.beta12,n);
119 end
120 pu=normcdf((Su-muu)/sdu);
121 pf=normcdf((Sf-muf)/sdf);
122 else %using exact distribution of Su, Sf
123 load('Matlab_SaveFiles/GRaPQuantile.mat',strcat('quant',num2str(k)));
124 switch k
125     case 1; quant=quant1;
126     case 2; quant=quant2;
127     case 3; quant=quant3;
128     case 4; quant=quant4;
129 end
130 pu=sum(quant(n,:,1)<=Su)/size(quant,2);
131 pf=sum(quant(n,:,2)<=Sf)/size(quant,2);
132 end
133 end
134
135 %plot of U and F
136 function property1(n,k)
137 hold on
138 set(gcf, 'PaperSize',[15 10], 'position',[100 100 600 400], 'color',[1 1 1]);
139 set(gca, 'xlim',[1-k n+k], 'ylim',[1-k n+k], 'box','on')
140 title(' \fontsize{14}dataset ', 'FontWeight','bold');
141 xlabel(' \fontsize{14}rank(x) ');
142 ylabel(' \fontsize{14}rank(y) ');
143 end

```

3 Vergleich der Unabhängigkeitstests

3.1 `IndepTestTool` - eine neue Toolbox für Matlab

Zur Beurteilung, welcher Test für welche Art von Zusammenhängen am Besten geeignet ist, sollten wir mehrere Tests gleichzeitig auf einen gegebenen Datensatz anwenden. So lassen sich durch Empirie schnell Gemeinsamkeiten und Unterschiede aufdecken. Da wir nun alle behandelten Tests in programmierter Form vorliegen haben, ist es nicht schwierig eine grafische Benutzeroberfläche zu erstellen. In interaktiver Weise sollen dem Nutzer verschiedene grafische Auswertungen eines Datensatzes ermöglicht werden, als auch eine zeitnahe Ausgabe sämtlicher p -Werte der behandelten Unabhängigkeitstests. Ein Anfang zur Auswertung bivariater Daten soll `IndepTestTool` darstellen. Das neue Matlab-Tool finden wir auf der beigelegten CD im Ordner `Matlab` und kann nach Wechsel des aktuellen Verzeichnisses per Befehlszeile in Matlab aufgerufen werden (`>> IndepTestTool`). Es öffnet sich dann die Benutzeroberfläche mit dem Fliegendatensatz. Wir können nun durch Nutzung der Klickboxen, Buttons und Scrollbalken die Darstellung des Plots ändern, Parameter der Unabhängigkeitstests umstellen oder zu anderen Datensätzen wechseln. So ist beispielsweise ein Wechsel zwischen dem Streudigramm der Originaldaten und dem Streudiagramm der Ränge möglich. Wir können die Punkt- und Flächenüberdeckungen des GRaP-Tests sowie die längste aufsteigende Teilfolge anzeigen lassen. Bei Datensätzen, die Zufallspunkte generieren, können wir die Anzahl der Punkte bestimmen, entweder durch Betätigung des zugehörigen Scrollbalkens, oder durch Eingabe in einem Textfeld. Bei einigen Datensätzen, wie `Linear2` ist zudem noch die Möglichkeit implementiert, die Varianz der Daten zu verändern oder die Daten zu rotieren. Parameter der Unabhängigkeitstests, wie die Bindungsdimension des Permutation-Entropy-Testes und der Radius der Umgebungen des GRaP Independence Tests können ebenfalls variiert werden. Natürlich besteht auch die Möglichkeit, Daten selbst einzugeben. Dazu schalten wir auf „Data Input“, geben die Variablen x und y ein und bestätigen mit „calculate“. Die Toolbox erlaubt dabei Eingaben von 6 bis 100 Datenpaaren und benötigt zur Auswertung und Anzeige auf Dual-Core-Prozessoren durchschnittlich nur eine Sekunde. Für $n=80$ rund zwei Sekunden, bei kleineren Datensätzen, wie für $n=25$, nur eine halbe Sekunde.

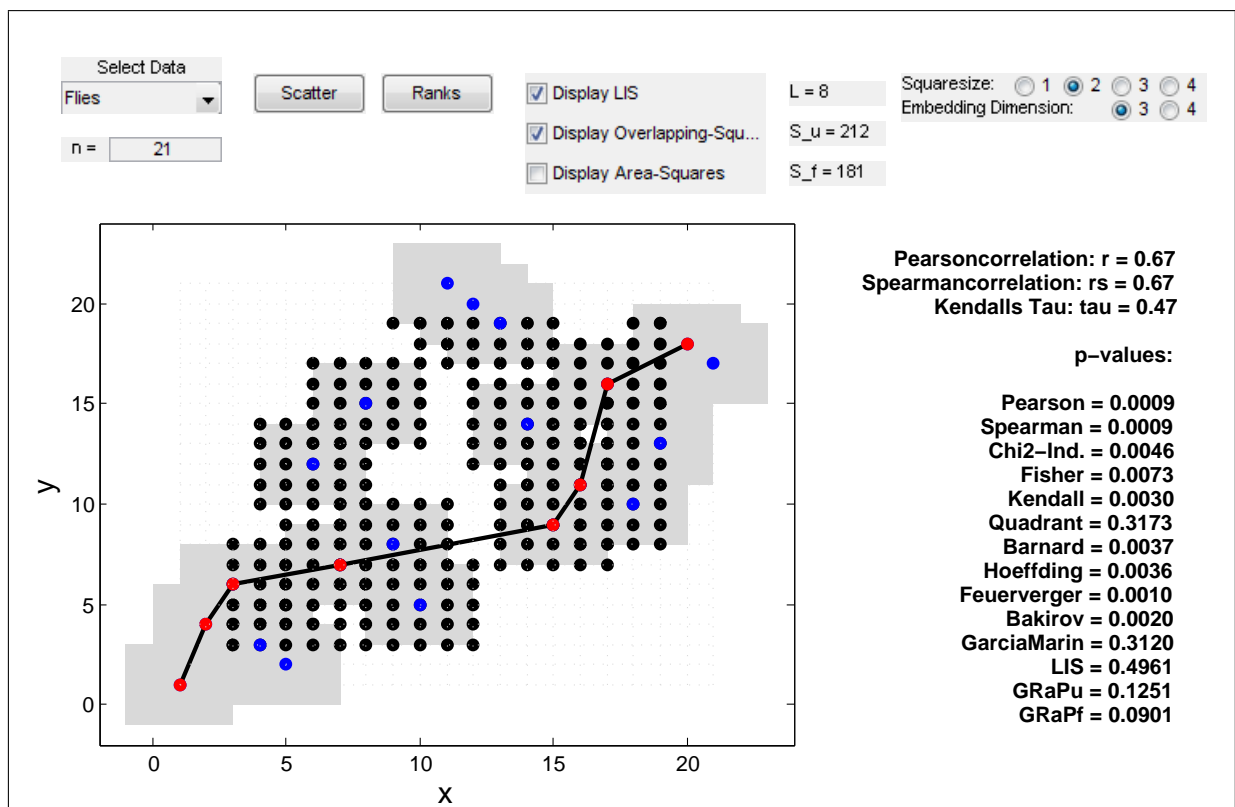


Abbildung 3.1: Benutzeroberfläche des *IndepTestTool*. Fliegendatensatz mit Darstellung der Punktüberdeckungen und längster aufsteigenden Teilfolge.

Wir haben somit ein Tool zur schnellen Überprüfung der Unabhängigkeit eigener Daten und zum Verhalten der Tests in unterschiedlichen theoretischen Situationen geschaffen. Eine genauere wissenschaftliche Auswertung zur Wirksamkeit der Tests bleibt uns aber nicht erspart und ist Thema des nächsten Abschnitts.

3.2 Power Analyse

Die Güte eines Tests richtet sich danach, ob mit dem Test die Nullhypothese, dass die Merkmale X und Y unabhängig sind, dann und nur dann abgelehnt wird, wenn in dem Datensatz eine Abhängigkeitsstruktur vorhanden ist. Ein guter Test würde also im unabhängigen Fall keine oder nur geringe Signifikanz ausgeben, während im abhängigen Fall eine starke Signifikanz wünschenswert ist. Im unabhängigen Fall interessiert uns daher die Wahrscheinlichkeit, die Nullhypothese anzunehmen (Spezifität): $P(H_0 \text{ akzeptiert} | H_0 \text{ gilt})$. Die Gegenwahrscheinlichkeit entspricht dem Fehler 1. Art (α -Fehler). Im abhängigen Fall interessiert uns die Wahrscheinlichkeit, die Nullhypothese abzulehnen: $P(H_1 | H_1)$. Dieser Prozentsatz wird auch mit Teststärke, Power eines Testes oder Sensitivität bezeichnet. Deren Gegenwahrscheinlichkeit mit Fehler 2. Art (β -Fehler). Im Folgenden wollen wir unsere Unabhängigkeitstests bei verschiedenen Unabhängigkeits- und Abhängigkeitsstrukturen auf Spezifität und Sensitivität überprüfen. Für jede Struktur werden wir zur Schätzung der Teststärke bzw. Spezifität je 10000 Stichproben für $n=10, 20, \dots, 100$ simulieren und die p-Werte berechnen. Für den GRaP Independence Test wählen wir für die verschiedenen n jeweils das empfohlene k . Wir werden im Folgenden auf die Werte vom χ^2 -, Fisher-, Quadranten- und Barnardtest verzichten, da die starre Klasseneinteilung über die Mediane keine gute Ergebnisse liefert und auch in der Praxis nicht angewendet werden sollte.

3.2.1 Der unabhängige Fall

Zum unabhängigen Fall wollen wir nur zwei Situationen untersuchen. Als Erstes testen wir die Spezifität bei der bivariaten Gleichverteilung.

Teststärke $P(1-\beta)$ für unabhängige bivariate gleichverteilte Zufallsgrößen										
Test	10	20	30	40	50	60	70	80	90	100
Pearson	0.946	0.948	0.948	0.952	0.948	0.949	0.951	0.954	0.954	0.952
Spearman	0.950	0.949	0.949	0.952	0.948	0.947	0.951	0.953	0.954	0.951
Kendall	0.952	0.951	0.949	0.952	0.950	0.950	0.953	0.953	0.955	0.952
Hoeffding	0.949	0.947	0.948	0.951	0.950	0.948	0.952	0.950	0.954	0.950
Bakirov	0.927	0.938	0.950	0.950	0.952	0.944	0.953	0.952	0.951	0.951
Garcia/Marin	0.949	0.954	0.955	0.950	0.951	0.952	0.953	0.951	0.956	0.952
LIS	0.986	0.982	0.989	0.977	0.975	0.985	0.984	0.978	0.974	0.972
GRaPu	0.958	0.952	0.954	0.952	0.953	0.951	0.950	0.949	0.950	0.953
GRaPf	0.944	0.952	0.956	0.953	0.952	0.951	0.950	0.950	0.948	0.956

Es zeigt sich, dass der Test über die längste aufsteigende Teilfolge (LIS) eine höhere Spezifität aufweist, als die anderen Tests, welche weitestgehend das Alpha-Niveau von 5% einhalten. Nur Bakirov zeigt Auffälligkeiten für kleinere n .

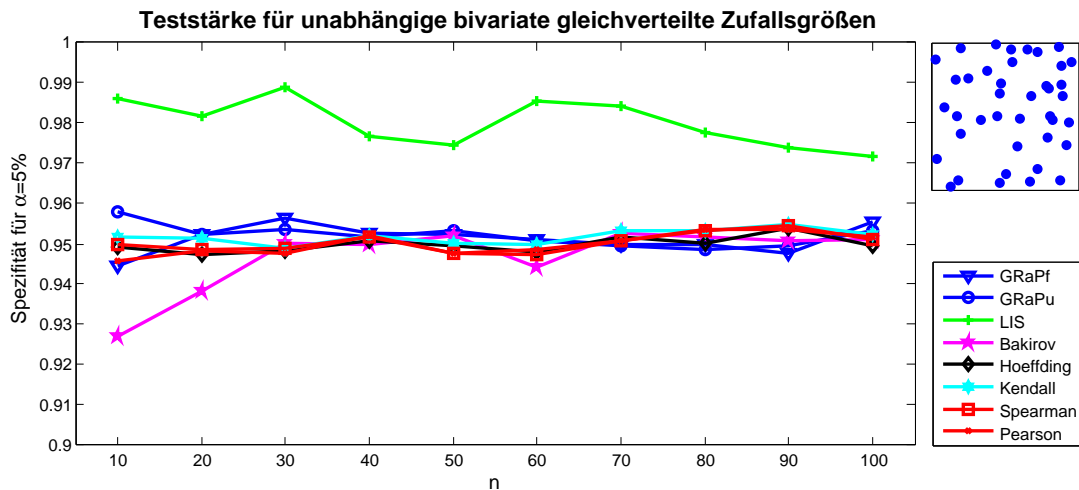


Abbildung 3.2: Stichprobenumfang vs. Spezifität für $\alpha=5\%$.

Abbildung 3.3 zeigt uns als Zweites die Spezifität der Tests für die bivariate Normalverteilung. Auch hier halten alle angewendeten Tests das Alpha-Niveau von 5% ein. Der LIS-Test von Garcia/Marin zeigt eine ähnlich verlaufende Kurve, wie beim unabhängigen gleichverteilten Fall. Die Kurve von Bakirovs Test liegt beim normalverteilten Fall aber sichtlich höher.

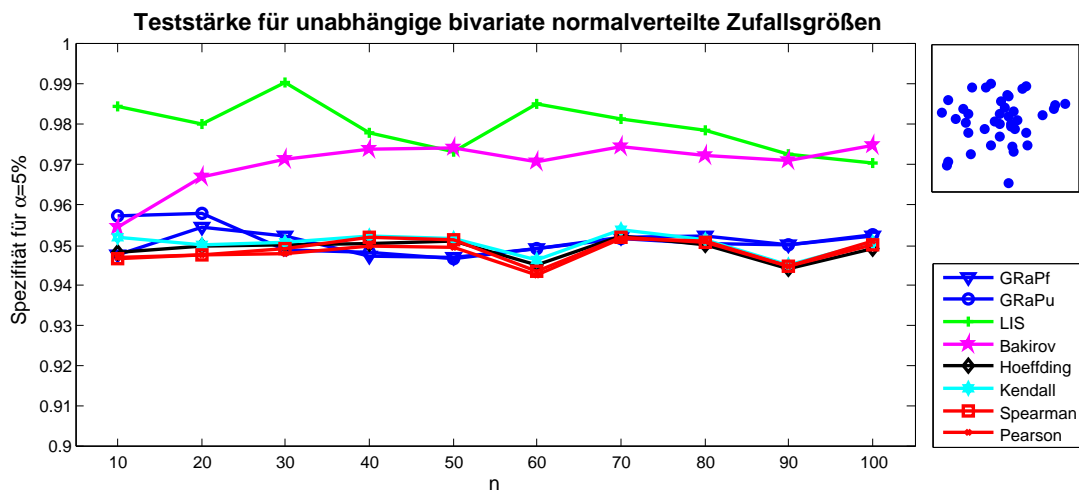


Abbildung 3.3: Stichprobenumfang vs. Spezifität für $\alpha=5\%$.

3 Vergleich der Unabhängigkeitstests

Teststärke $P(1-\beta)$ für unabhängige bivariate normalverteilte Zufallsgrößen										
Test	10	20	30	40	50	60	70	80	90	100
Pearson	0.947	0.948	0.948	0.950	0.950	0.943	0.952	0.951	0.945	0.951
Spearman	0.947	0.948	0.949	0.952	0.951	0.944	0.952	0.951	0.945	0.950
Kendall	0.952	0.950	0.951	0.952	0.952	0.947	0.954	0.951	0.945	0.951
Hoeffding	0.948	0.950	0.950	0.951	0.951	0.945	0.952	0.950	0.944	0.949
Bakirov	0.955	0.967	0.971	0.974	0.974	0.971	0.974	0.972	0.971	0.975
Garcia/Marin	0.948	0.953	0.957	0.955	0.950	0.955	0.950	0.953	0.955	0.952
LIS	0.985	0.980	0.991	0.978	0.973	0.985	0.981	0.979	0.973	0.971
GRaPu	0.957	0.958	0.949	0.948	0.947	0.949	0.952	0.950	0.950	0.953
GRaPf	0.948	0.955	0.952	0.947	0.947	0.949	0.952	0.952	0.950	0.952

3.2.2 Der abhängige Fall

Wir betrachten nun einige ausgewählte Strukturen zur Abhängigkeit betrachten. Beginnen wollen wir mit einem linearen Fall. Abbildung 3.4 zeigt uns die empirische Powerfunktion in Abhängigkeit von n . Wie zu erwarten ist die Teststärke von Pearson und Spearman hervorragend. Der LIS- und der GRaP-Test haben dagegen für kleinere n einige Probleme, die Abhängigkeit zu erkennen. Vor allem, wenn wir die Varianz entlang der Regressionsgeraden erhöhen, werden die p-Werte schlechter. Gleiche Anfangsschwierigkeiten gibt es auch für den exponentiellen Zusammenhang (Abbildung 3.5), der eine monotone Abhängigkeit repräsentiert.

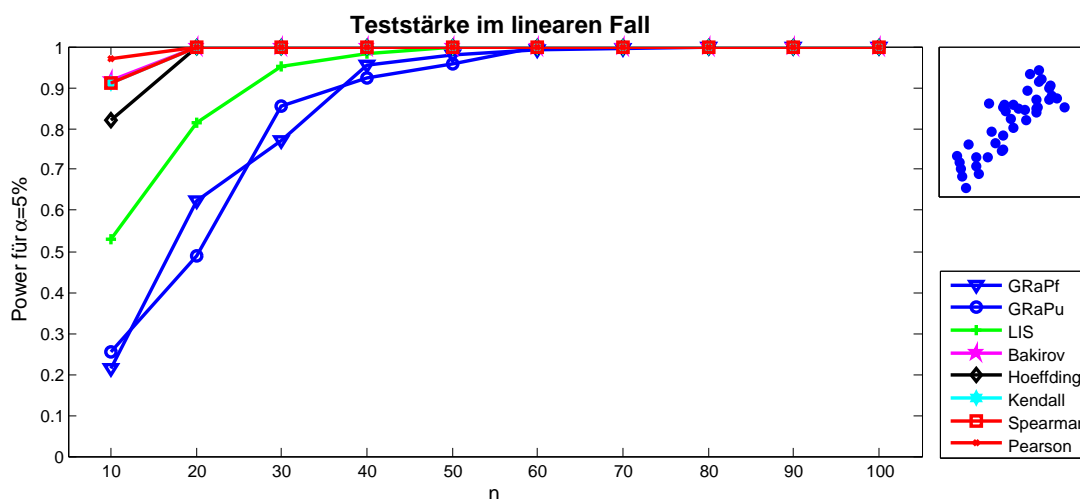


Abbildung 3.4: Stichprobenumfang vs. Teststärke für $\alpha=5\%$.

Teststärke $P(1-\beta)$ im linearen Fall										
Test	10	20	30	40	50	60	70	80	90	100
Pearson	0.971	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Spearman	0.912	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Kendall	0.910	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Hoeffding	0.820	0.997	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Bakirov	0.916	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Garcia/Marin	0.208	0.157	0.125	0.113	0.111	0.096	0.093	0.083	0.086	0.084
LIS	0.530	0.813	0.951	0.982	0.998	0.998	1.000	1.000	1.000	1.000
GRaPu	0.255	0.490	0.853	0.923	0.959	0.997	0.998	0.999	1.000	1.000
GRaPf	0.216	0.624	0.771	0.954	0.981	0.992	0.995	1.000	1.000	1.000

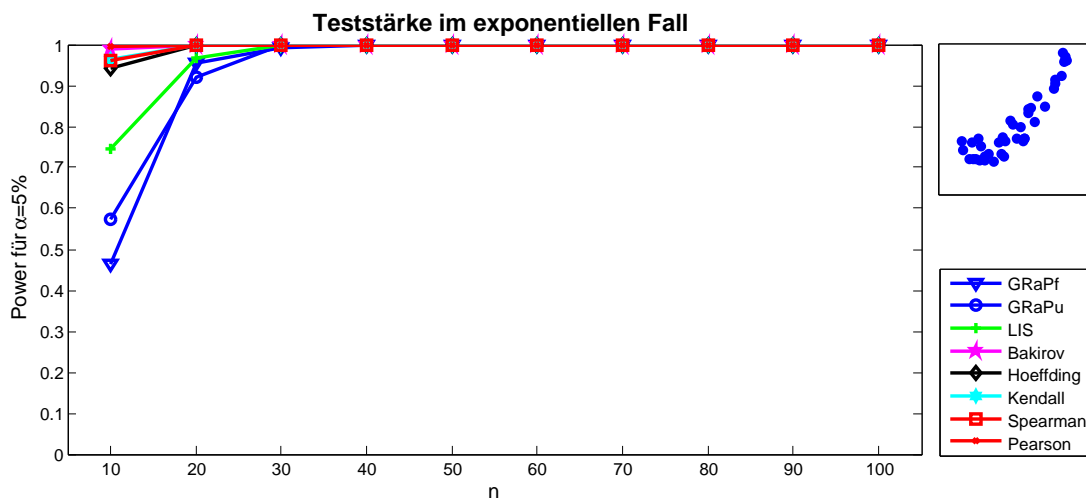


Abbildung 3.5: Stichprobenumfang vs. Teststärke für $\alpha=5\%$.

Teststärke $P(1-\beta)$ im exponentiellen Fall										
Test	10	20	30	40	50	60	70	80	90	100
Pearson	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Spearman	0.962	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Kendall	0.964	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Hoeffding	0.942	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Bakirov	0.988	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Garcia/Marin	0.410	0.391	0.331	0.308	0.276	0.235	0.197	0.168	0.154	0.149
LIS	0.744	0.969	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GRaPu	0.575	0.920	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
GRaPf	0.464	0.955	0.993	1.000	1.000	1.000	1.000	1.000	1.000	1.000

3 Vergleich der Unabhängigkeitstests

Als nächstes betrachten wir einen doppelten linearen Zusammenhang. Damit ist gemeint, dass wir im Streudiagramm lineare Zusammenhänge in zwei Komponenten erkennen können. Dieser Fall kann insbesondere dann auftreten, wenn ein Datensatz in 2 Klassen zerfällt. Gute Ergebnisse erzielen die Tests von Hoeffding und Bakirov. Um eine Klasse besser scheint jedoch der GRaP Independence Test zu sein. Selbst bei einer geringen Stichprobengröße von $n=30$ sind, bei einem Alpha-Niveau von 5%, bereits 60% der p-Werte signifikant.

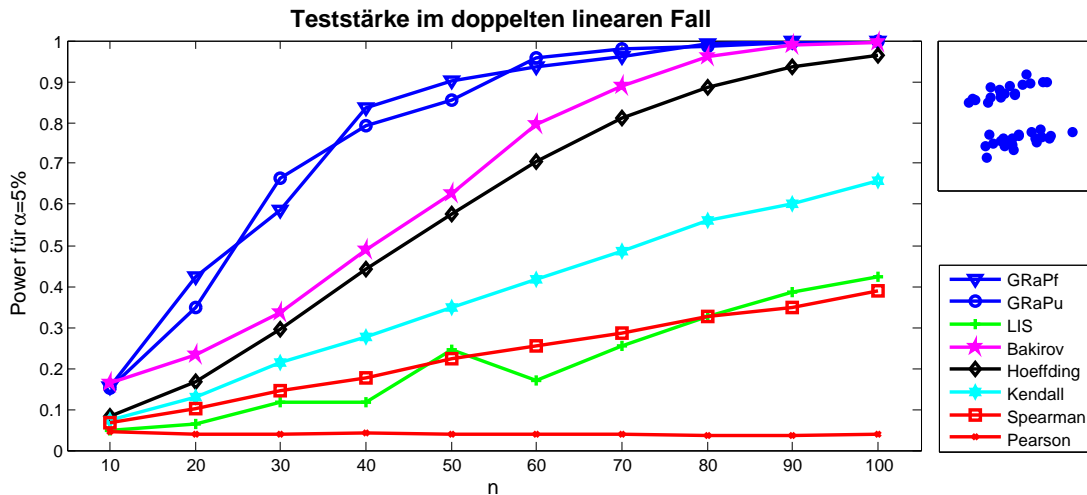


Abbildung 3.6: Stichprobenumfang vs. Teststärke für $\alpha=5\%$.

Teststärke $P(1-\beta)$ im doppelten linearen Fall										
Test	10	20	30	40	50	60	70	80	90	100
Pearson	0.044	0.038	0.039	0.043	0.040	0.041	0.040	0.035	0.036	0.038
Spearman	0.066	0.100	0.146	0.176	0.222	0.253	0.286	0.325	0.349	0.389
Kendall	0.075	0.129	0.214	0.278	0.350	0.419	0.486	0.560	0.603	0.658
Hoeffding	0.084	0.167	0.295	0.441	0.576	0.704	0.811	0.887	0.936	0.966
Bakirov	0.164	0.232	0.337	0.489	0.626	0.796	0.890	0.960	0.989	0.996
Garcia/Marin	0.055	0.052	0.051	0.050	0.055	0.052	0.053	0.048	0.050	0.050
LIS	0.048	0.065	0.117	0.117	0.247	0.169	0.255	0.326	0.387	0.424
GRaPu	0.152	0.349	0.663	0.792	0.854	0.958	0.979	0.987	0.994	0.999
GRaPf	0.156	0.423	0.587	0.837	0.901	0.937	0.961	0.993	0.997	0.998

Das nächste Beispiel war schon Bestandteil der Veröffentlichung von García und González-López [45]. Es handelt sich um einen Mix aus zwei bivariaten Normalverteilungen, deren erste Verteilung standardnormalverteilt ist mit einer Korrelation ρ , während die andere standardnormalverteilt ist mit einem Korrelationskoeffizienten von $-\rho$. Die Simulation

wurde für unterschiedliche ρ durchgeführt. Der Kurvenverlauf der längsten aufsteigenden Teilfolge aus [45] kann bestätigt werden. Zudem zeigt sich, dass der GRaP Independence Test in seiner Teststärke konkurrenzfähig ist. Bakirov jedoch erkennt derartige Zusammenhänge mit Abstand am Besten.

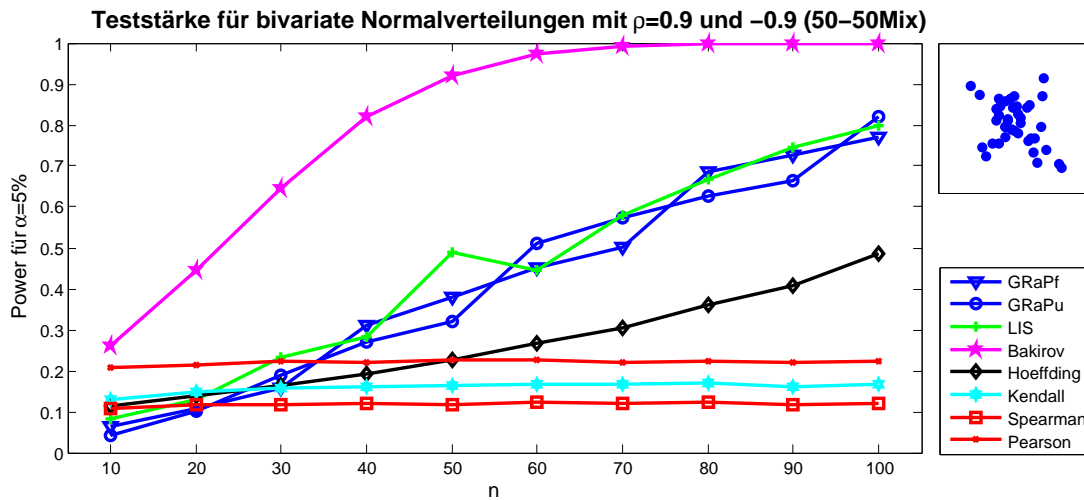


Abbildung 3.7: Stichprobenumfang vs. Teststärke für $\alpha=5\%$.

Teststärke $P(1-\beta)$ im Mix zweier bivariater Normalverteilungen											
ρ	Test	10	20	30	40	50	60	70	80	90	100
0.9	Pearson	0.208	0.215	0.224	0.221	0.227	0.226	0.222	0.223	0.220	0.223
	Spearman	0.108	0.116	0.117	0.120	0.118	0.125	0.120	0.124	0.117	0.121
	Kendall	0.130	0.148	0.158	0.160	0.164	0.168	0.168	0.170	0.161	0.167
	Hoeffding	0.113	0.139	0.165	0.194	0.227	0.268	0.304	0.360	0.408	0.487
	Bakirov	0.260	0.444	0.644	0.821	0.920	0.975	0.993	0.998	1.000	1.000
	Garcia/Marin	0.056	0.052	0.045	0.044	0.049	0.046	0.051	0.044	0.047	0.047
	LIS	0.081	0.129	0.233	0.284	0.488	0.445	0.581	0.666	0.747	0.798
	GRaPu	0.042	0.102	0.190	0.271	0.321	0.512	0.573	0.626	0.665	0.821
	GRaPf	0.063	0.109	0.158	0.310	0.379	0.450	0.500	0.687	0.728	0.771
0.8	Pearson	0.169	0.184	0.187	0.187	0.192	0.191	0.191	0.195	0.188	0.191
	Spearman	0.091	0.101	0.097	0.101	0.102	0.106	0.105	0.108	0.103	0.103
	Kendall	0.106	0.120	0.126	0.131	0.131	0.139	0.132	0.138	0.131	0.131
	Hoeffding	0.086	0.107	0.110	0.126	0.139	0.153	0.163	0.177	0.178	0.201
	Bakirov	0.156	0.215	0.285	0.408	0.499	0.634	0.705	0.808	0.859	0.910
	Garcia/Marin	0.049	0.053	0.046	0.043	0.047	0.046	0.047	0.046	0.047	0.049
	LIS	0.050	0.066	0.109	0.119	0.240	0.185	0.270	0.329	0.392	0.444
	GRaPu	0.031	0.049	0.063	0.084	0.096	0.132	0.148	0.169	0.185	0.243
	GRaPf	0.050	0.051	0.061	0.093	0.102	0.122	0.131	0.185	0.197	0.220

3 Vergleich der Unabhängigkeitstests

Teststärke $P(1-\beta)$ im Mix zweier bivariater Normalverteilungen											
ρ	Test	10	20	30	40	50	60	70	80	90	100
0.7	Pearson	0.134	0.153	0.157	0.155	0.159	0.162	0.162	0.165	0.156	0.164
	Spearman	0.081	0.089	0.083	0.087	0.085	0.091	0.088	0.093	0.089	0.088
	Kendall	0.089	0.102	0.101	0.106	0.106	0.109	0.108	0.114	0.110	0.107
	Hoeffding	0.073	0.088	0.088	0.094	0.098	0.108	0.111	0.117	0.116	0.123
	Bakirov	0.106	0.118	0.141	0.185	0.213	0.280	0.315	0.396	0.440	0.502
	Garcia/Marin	0.047	0.048	0.043	0.045	0.049	0.049	0.048	0.046	0.049	0.050
	LIS	0.038	0.043	0.060	0.066	0.131	0.084	0.133	0.163	0.197	0.232
	GRaPu	0.030	0.041	0.039	0.053	0.052	0.061	0.066	0.069	0.078	0.088
	GRaPf	0.045	0.038	0.039	0.051	0.055	0.059	0.062	0.072	0.079	0.087
0.6	Pearson	0.107	0.121	0.129	0.125	0.132	0.133	0.134	0.138	0.132	0.135
	Spearman	0.070	0.078	0.074	0.075	0.075	0.078	0.078	0.081	0.078	0.080
	Kendall	0.076	0.084	0.085	0.087	0.090	0.089	0.089	0.096	0.093	0.093
	Hoeffding	0.067	0.074	0.072	0.076	0.081	0.085	0.084	0.091	0.088	0.092
	Bakirov	0.080	0.077	0.082	0.094	0.102	0.128	0.130	0.164	0.175	0.200
	Garcia/Marin	0.047	0.049	0.041	0.046	0.049	0.047	0.046	0.044	0.048	0.049
	LIS	0.031	0.031	0.040	0.039	0.080	0.046	0.071	0.094	0.111	0.123
	GRaPu	0.036	0.037	0.036	0.039	0.043	0.043	0.042	0.046	0.049	0.050
	GRaPf	0.050	0.034	0.037	0.039	0.043	0.042	0.041	0.045	0.050	0.049

Die letzten beiden Fälle, die wir uns anschauen wollen, sind ein sinusförmiger und ein kreisförmiger Zusammenhang. Beide Abhängigkeiten wurden jeweils durch Addition normalverteilter Zufallsgrößen mit Mittelwert 0 entlang der Zusammenhangskurve verstreut. Die Abhängigkeiten wurden dabei jeweils mit zwei unterschiedlichen Varianzen getestet. Abbildung 3.8 und 3.9 zeigen ein sehr überzeugendes Ergebnis für den GRaP Independence Test, dessen Power beim kreisförmigen Zusammenhang sogar Bakirov und Hoeffding übertrifft.

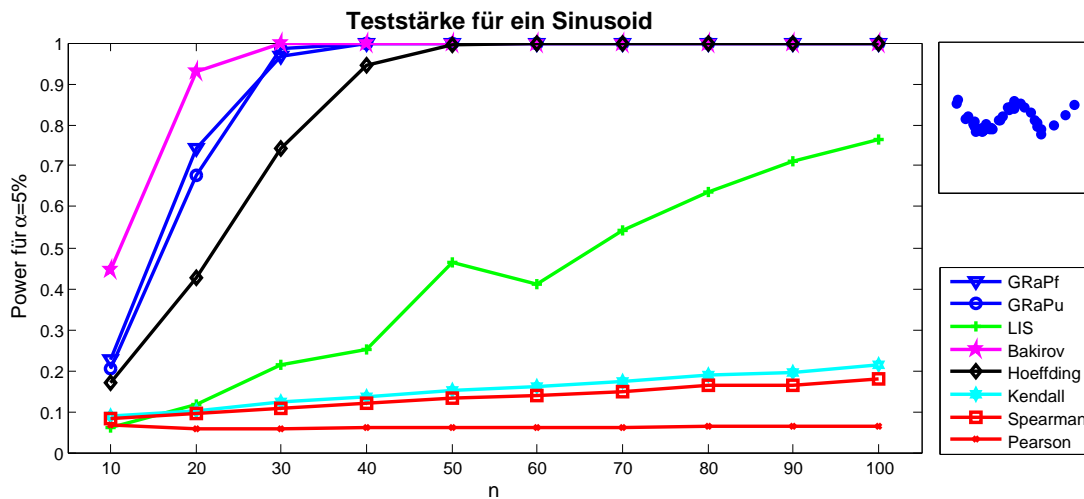


Abbildung 3.8: Stichprobenumfang vs. Teststärke für $\alpha=5\%$.

Teststärke $P(1-\beta)$ für ein Sinusoid											
Var	Test	10	20	30	40	50	60	70	80	90	100
0.2	Pearson	0.068	0.057	0.057	0.061	0.062	0.060	0.060	0.066	0.063	0.064
	Spearman	0.083	0.094	0.107	0.119	0.132	0.139	0.148	0.163	0.164	0.180
	Kendall	0.089	0.102	0.123	0.136	0.151	0.161	0.172	0.189	0.196	0.214
	Hoeffding	0.171	0.426	0.743	0.945	0.995	1.000	1.000	1.000	1.000	1.000
	Bakirov	0.446	0.928	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Garcia/Marin	0.086	0.217	0.238	0.238	0.213	0.184	0.167	0.153	0.137	0.125
	LIS	0.060	0.118	0.213	0.253	0.464	0.412	0.542	0.636	0.712	0.764
	GRaPu	0.205	0.676	0.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	GRaPf	0.227	0.743	0.966	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.4	Pearson	0.062	0.054	0.057	0.056	0.058	0.058	0.060	0.058	0.060	0.060
	Spearman	0.077	0.085	0.092	0.102	0.113	0.126	0.138	0.138	0.146	0.157
	Kendall	0.076	0.085	0.097	0.109	0.117	0.132	0.146	0.149	0.159	0.169
	Hoeffding	0.126	0.260	0.457	0.681	0.860	0.961	0.993	0.999	1.000	1.000
	Bakirov	0.284	0.691	0.942	0.996	1.000	1.000	1.000	1.000	1.000	1.000
	Garcia/Marin	0.064	0.089	0.082	0.074	0.079	0.065	0.058	0.062	0.055	0.056
	LIS	0.039	0.051	0.077	0.082	0.165	0.117	0.178	0.224	0.258	0.295
	GRaPu	0.145	0.421	0.832	0.945	0.974	1.000	1.000	1.000	1.000	1.000
	GRaPf	0.158	0.511	0.740	0.971	0.993	0.998	0.999	1.000	1.000	1.000

3 Vergleich der Unabhängigkeitstests

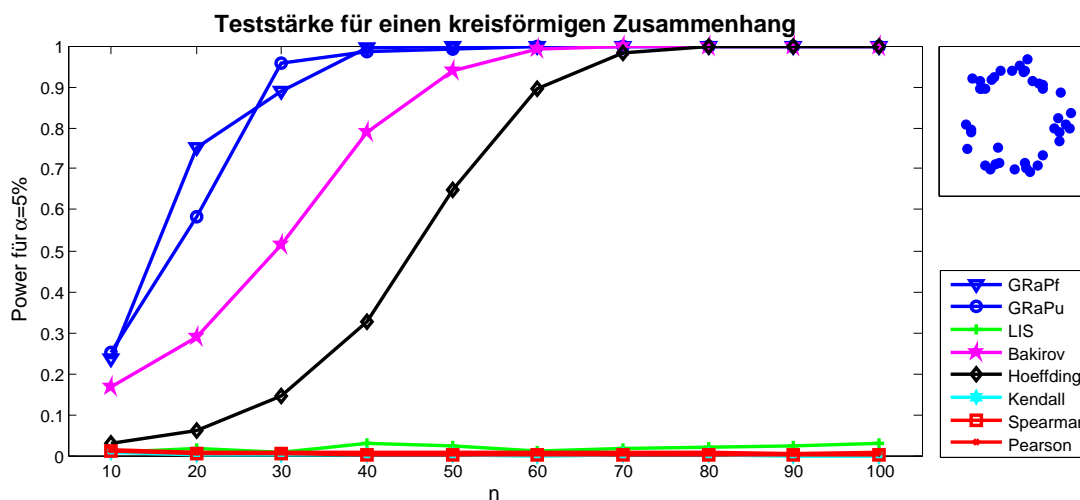


Abbildung 3.9: Stichprobenumfang vs. Teststärke für $\alpha=5\%$.

Teststärke $P(1-\beta)$ für einen kreisförmigen Zusammenhang											
Var	Test	10	20	30	40	50	60	70	80	90	100
0.1	Pearson	0.016	0.009	0.007	0.008	0.006	0.006	0.007	0.008	0.005	0.006
	Spearman	0.010	0.006	0.003	0.002	0.002	0.003	0.002	0.002	0.001	0.002
	Kendall	0.009	0.003	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Hoeffding	0.030	0.062	0.144	0.326	0.647	0.895	0.984	0.998	1.000	1.000
	Bakirov	0.167	0.288	0.515	0.788	0.939	0.993	0.998	1.000	1.000	1.000
	Garcia/Marin	0.039	0.046	0.049	0.049	0.049	0.048	0.048	0.047	0.051	0.048
	LIS	0.009	0.018	0.006	0.029	0.024	0.013	0.016	0.021	0.022	0.028
	GRaPu	0.251	0.582	0.958	0.987	0.993	1.000	1.000	1.000	1.000	1.000
	GRaPf	0.236	0.752	0.888	0.996	0.999	1.000	1.000	1.000	1.000	1.000
0.2	Pearson	0.017	0.013	0.011	0.011	0.008	0.009	0.011	0.011	0.009	0.010
	Spearman	0.014	0.011	0.009	0.010	0.007	0.008	0.008	0.009	0.007	0.007
	Kendall	0.011	0.004	0.003	0.002	0.001	0.002	0.002	0.002	0.001	0.001
	Hoeffding	0.026	0.028	0.035	0.052	0.079	0.119	0.179	0.257	0.357	0.473
	Bakirov	0.102	0.130	0.199	0.322	0.439	0.621	0.731	0.854	0.925	0.962
	Garcia/Marin	0.041	0.048	0.045	0.048	0.051	0.049	0.050	0.046	0.049	0.047
	LIS	0.008	0.015	0.004	0.040	0.029	0.027	0.028	0.035	0.047	0.070
	GRaPu	0.146	0.244	0.500	0.570	0.581	0.812	0.825	0.849	0.859	0.954
	GRaPf	0.139	0.337	0.384	0.665	0.693	0.726	0.740	0.901	0.914	0.918

3.3 Praktische Beispiele

Den Fliegendatensatz konnten wir begleitend zur Theorie betrachten. Auch verschiedenste Konstruktionen von Zusammenhängen haben wir bereits in der Power-Analyse kennengelernt. Abschließend betrachten wir noch einige Datensätze aus der Praxis. Beginnen wollen wir mit dem Iris Datensatz von Fisher.

Fishers Iris Datensatz

Der multivariate Datensatz enthält die Längen und Breiten von den Kelch- und Kronblättern von je 50 Schwertlilien der Gattungen *Setosa*, *Versicolor* und *Virginica*. Er ist einer der bekanntesten frei erhältlichen Datensätze und kann über `load fisheriris` in Matlab geladen werden. Wir betrachten die Breite der Kelch- und Kronblätter. Wir zeichnen das verzitterte Streudiagramm. Nahezu alle bekannten Tests haben einen signifikanten

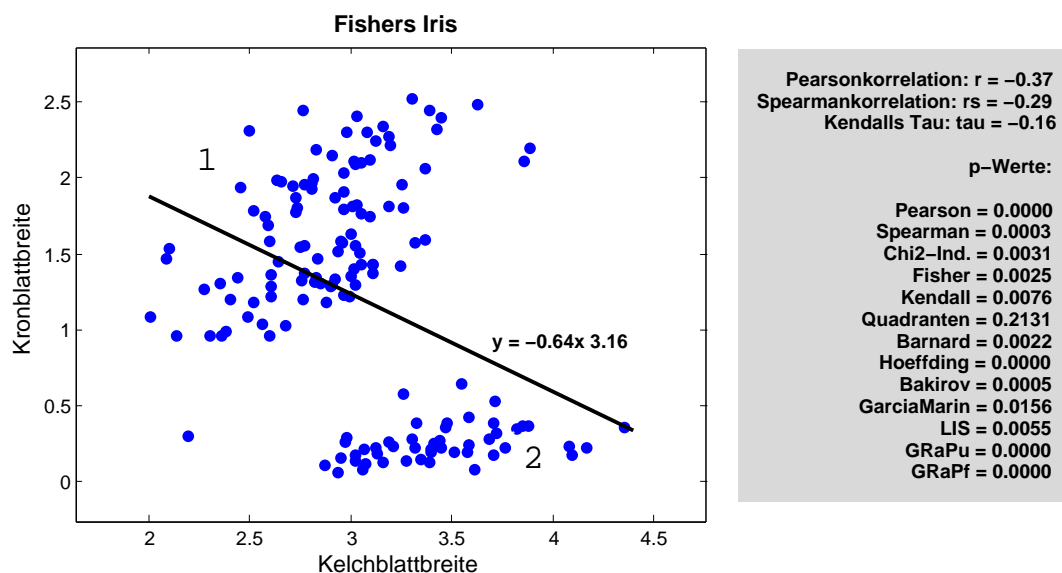


Abbildung 3.10: Verzittertes Streudiagramm und Regressionsgerade des Irisdatensatzes sowie die p-Werte verschiedener Unabhängigkeitstests. Die drei Weinsorten jeweils in Blau, Rot und Grün

p-Wert. Eine Abhängigkeit zwischen der Kelchblattbreite und Kronblattbreite ist somit auch statistisch nachgewiesen. Schauen wir uns allerdings die Korrelationswerte von Pearson und Spearman an, so würden wir schließen, dass Kelchblattbreite und Kronblattbreite negativ korreliert sind. Die Natur zeigt uns genau das Gegenteil – es gibt einen positiven Zusammenhang! Grund für dieses Versagen liegt an der Anpassung der Daten an einer

3 Vergleich der Unabhängigkeitstests

Regressionsgeraden. Die Punktwolke 2, um eine Kelchblattbreite von 0.2 cm angeordnet, zieht die Gerade zweifellos ins Negative. Trotzdem werden die Daten fälschlicherweise gut angepasst. Daher sind die p-Werte auch signifikant. In Wirklichkeit aber haben nur die Gattungen *Versicolor* und *Virginica* einen Zusammenhang (Punktwolke 1). Die Gattung *Setosa* (Punktwolke 2) dagegen keinen. Ein Blick auf das Streudiagramm sollte deshalb nie erspart bleiben!

Chemische Analyse von Weinen

Stefan Aeberhard stellte dem Archiv des UCI seinerseits einen Datensatz zur chemischen Analyse von Weinen zur Verfügung [53]. Der Datensatz umfasst 13 chemische Variablen von 178 Weinen einer Region Italiens. Es werden dabei 3 Sorten unterschieden (59 Weine der ersten Sorte, 71 Weine der zweiten und 48 der dritten Sorte). Neben dem Alkoholgehalt wurden auch der Gehalt an Apfelsäure, Magnesium und Asche gemessen. Wir wollen die Abhängigkeit zwischen den nichtflavonoiden Phenolen und dem Farbton des Weines betrachten. Wir erkennen einen positiven Zusammenhang zwischen der Konzen-

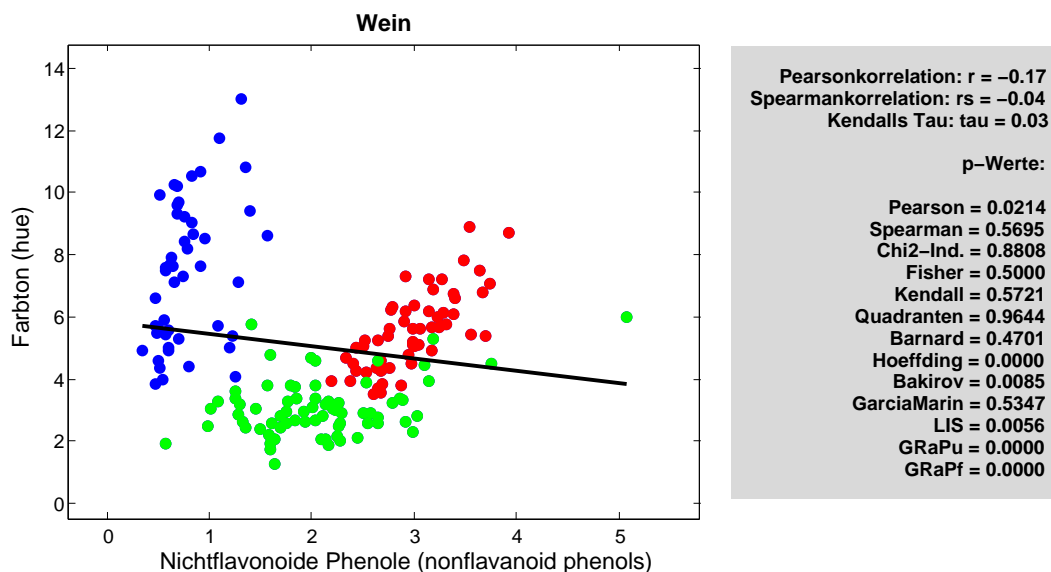


Abbildung 3.11: Zusammenhang zwischen nichtflavonoide Phenolen und Farbton des Weines mit Regressionsgeraden und p-Werten.

tration der nichtflavonoiden Phenolen und dem Farbton, zumindest bei den Sorten in Rot und Grün. Nach [40] zu urteilen sind eigentlich Flavonoide, wie Anthocyane und Tannine für die Farbgebung des Weines verantwortlich und nicht die Phenolcarbonsäuren (nicht-

flavonide Phenole). Die Erklärung ist recht einfach: Bei den roten und grünen Sorten steigt der Gehalt an Flavonoiden mit der Konzentration der Nicht-Flavonoide, während bei der blauen Sorte kein Zusammenhang bestätigt werden kann. Die statistischen Tests sollten demnach einen Zusammenhang, bedingt durch die Kausalität Flavonoide/Nicht-Flavonoide, signalisieren. Der Hoeffding-, Bakirov-, LIS- und GRaP-Test haben höchst

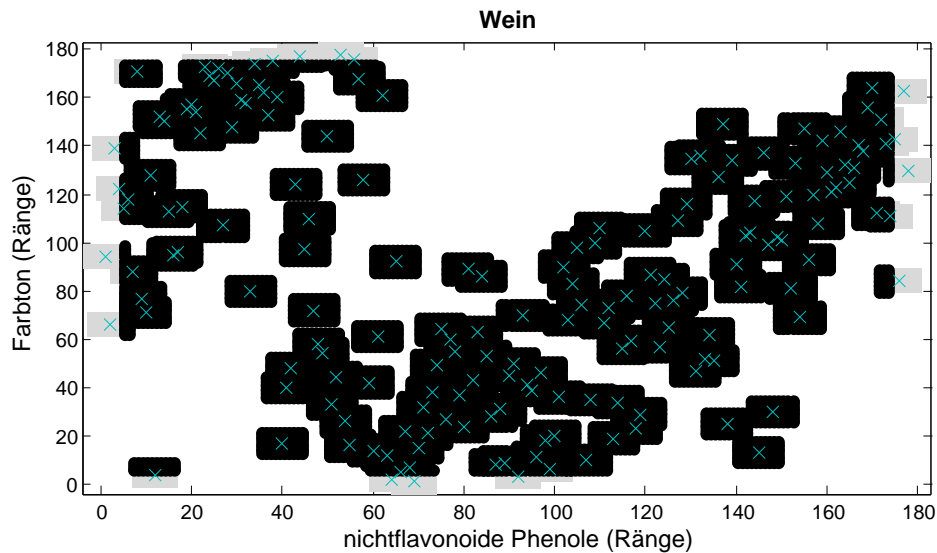


Abbildung 3.12: Signifikant wenig Flächenüberdeckungen zwischen nichtflavonoide Phenolen und dem Farbton von Weinen.

signifikante p-Werte und bestätigen die Abhängigkeit. Pearson liefert zwar auch Signifikanz, wie im Iris-Beispiel, jedoch für einen negativen Zusammenhang. Fairerweise sollte aber darauf hingewiesen werden, dass der Barnard- und der Chi-Quadrat-Test sicherlich mit einer anderen Klasseneinteilung besser abschneiden würden.

Korrelation 21

Auf Wikipedia finden wir eine schöne Übersicht zu den Korrelationskoeffizienten von 21 Beispielen (siehe http://en.wikipedia.org/wiki/File:Correlation_examples.png). Diese Schar an Beispielen haben wir in Matlab übertragen, und auf ihren Zusammenhang hin untersucht. Dabei ist zu bemerken, dass die Beispiele der Nummern 4 und 21 durch Konstruktion unabhängig sind, während die anderen eine Abhängigkeit aufweisen. Getestet wurden dabei alle 21 Muster der Abbildung 3.13 mit je 50 Datenpunkten. Anhand der zugehörigen p-Werte, welche in Abbildung 3.14 visualisiert wurden, erkennen wir klare Schwächen einiger Tests. Wir stellen fest, dass die klassischen Verfahren nicht in der Lage sind, komplexere Abhängigkeitsstrukturen zu erkennen. Auch der Permutationstest

3 Vergleich der Unabhängigkeitstests

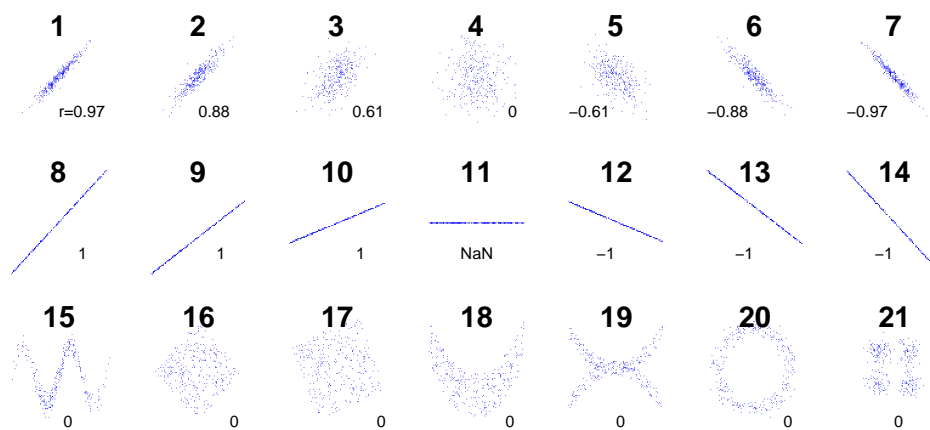


Abbildung 3.13: 21 Beispiele zur Pearsonkorrelation. Bildnummer mit zugehörigem Korrelationskoeffizienten nach Pearson.

von García/Marín und der Test über die längste aufsteigende Teilfolge von García und González-López helfen hier nicht. Wirklich überzeugen können die Tests von Hoeffding und Bakirov, sowie der neu vorgestellte Test über die Geometrie randomisierter Permutationen, auch wenn letzt genannter bei großen Streuungen einige Schwächen aufweist. So wird selbst ein linearer Zusammenhang mit größerer Streuung, wie in Bild Nummer 3 und 5, nicht erkannt.



Abbildung 3.14: Farblich gestufte p-Werte der Beispiele aus 3.13. Die numerischen Werte können dem Anhang entnommen werden (5.2).

4 Zusammenfassung

Stochastische Unabhängigkeit gehört zu den zentralen Begriffen der Wahrscheinlichkeitsrechnung und ist von grundlegender Bedeutung in der induktiven Statistik. Im praktischen Umgang mit Daten gehören Untersuchungen auf Zusammenhänge zweier Merkmale zum Tagesgeschäft vieler Wissenschaftler. Um so bedeutsamer ist es, einen wirksamen Test zu haben, dessen Teststärke nicht nur in einem linearen Zusammenhang besteht. Denn nur selten sind Abhängigkeiten linearer Natur und Alternativen zur Unabhängigkeit gibt es viele. Einen guten Test zeichnet eine hohe Teststärke (Power) in möglichst vielen Alternativen aus. Im Kapitel 2 haben wir verschiedene Unabhängigkeitstests kennen gelernt. Der GRaP Independence Test wurde dabei erstmalig detailliert beschrieben und zur Anwendung in Matlab aufbereitet. Der neue Unabhängigkeitstest hat eine leicht zu berechnende Teststatistik und eine Verteilungsfunktion mit guten Eigenschaften. Es entstand mit `IndepTestTool` eine neue Matlab-Toolbox, mit der sich leicht verschiedene Zusammenhänge konstruieren und testen lassen. Mit der Poweranalyse im Kapitel 3 haben wir dann einige abhängige Strukturen genauer untersucht. Dabei hat sich herausgestellt, dass der GRaP Independence Test vor allem im nichtlinearen Fall eine hohe Teststärke besitzt. GRaP stellt somit eine wahre Alternative zu den namenhaften Tests von Hoeffding und Bakirov dar.

Interessante Ausblicke

Es wäre sicherlich hilfreich, das `IndepTestTool` auf mehr als 100 Datenpunkte auszubauen. Der Import eigener Daten sollte ermöglicht werden und eine leichtere Anpassung der Daten wäre denkbar. So könnte beispielsweise das Hinzufügen und Entfernen eines Datenpunktes durch Mausklicks oder das Verschieben eines Punktes per Drag & Drop im Koordinatensystem ermöglicht werden. Ich brauche auch nicht zu verschweigen, dass noch andere interessante Tests, wie der Feuerverger Test [35], zu untersuchen wären. Währenddessen wird eine wissenschaftliche Publikation des GRaP Independence Tests angestrebt und der multivariate Ausbau forciert.
















Bibliographie

Bücher

- [1] M. G. KENDALL (1955), *Rank correlation methods*, Hafner Publications, New York.
- [2] M. G. KENDALL, J. D. GIBBONS (1990), *Rank correlation methods. 5th ed.*, Edward Arnold, London.
- [3] H. JOE (1997), *Multivariate models and dependence concepts*, Chapman & Hall/-CRC, London; New York.
- [4] D. D. MARI, S. KOTZ (2001), *Correlation and Dependence*, Imperial College Press, London.
- [5] L. FAHRMEIR, R. KÜNSTLER, I. PIGEOT, G. TUTZ(2004) *Statistik - Der Weg zur Datenanalyse, fünfte Auflage* Springer-Verlag, Berlin; Heidelberg.
- [6] J. BORTZ, G. A. LIENERT, K. BOEHNKE(2008) *Verteilungsfreie Methoden in der Biostatistik, 3. Auflage* Springer Medizin Verlag, Heidelberg.








Wissenschaftliche Publikationen

- [7] T. BAYES, R. PRICE (1763), *An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.*, Philosophical Transactions of the Royal Society of London, Vol. 53, pp 370-418. ☐
- [8] F. GALTON (1886), *Regression Towards Mediocrity in Hereditary Stature*, Journal of the Anthropological Institute, Vol. 15, pp. 246-263. ☐
- [9] K. PEARSON (1895), *Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material*, Philosophical Transactions of the Royal Society of London. A, Vol. 186, pp. 343-414. ☐
- [10] C. SPEARMAN (1904), *The Proof and Measurement of Association between Two Things*, The American Journal of Psychology, Vol. 15, pp. 72-101. ☐
- [11] K. PEARSON (1920), *Notes on the History of Correlation*, Biometrika, Vol. 13, pp. 25-45. ☐

- [12] K. PEARSON (1922), *On the χ^2 Test of Goodness of Fit*, Biometrika, Vol. 14, pp. 186-191. 
- [13] R. A. FISHER (1922), *On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P*, Journal of the Royal Statistical Society, Vol. 85, pp. 87-94. 
- [14] R. A. FISHER (1922), *On the Mathematical Foundations of Theoretical Statistics*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 222, pp. 309-368. 
- [15] M. G. KENDALL (1938), *A New Measure Of Rank Correlation*, Biometrika, Vol. 30, pp. 81-93. 
- [16] G. A. BARNARD (1945), *A New Test for 2×2 Tables*, Nature, Vol. 156, pp. 177-177. 
- [17] G. A. BARNARD (1947), *Significance Tests for 2×2 Tables*, Biometrika, Vol. 34, pp. 123-138. 
- [18] W. HOEFFDING (1948), *A Non-Parametric Test of Independence*, The Annals of Mathematical Statistics, Vol. 19, pp. 546-557. 
- [19] G. H. FREEMAN, J. H. HALTON (1951), *Note On An Exact Treatment Of Contingency, Goodness Of Fit And Other Problems Of Significance*, Biometrika, Vol. 38, pp. 141-149. 
- [20] J. S. FRAME, G. DE B. ROBINSON AND R.M. THRALL (1954), *The hook graphs of the symmetric group*, Canadian Journal of Mathematics, Vol. 6, pp. 316-324. 
- [21] J. R. BLUM, J. KIEFER AND M. ROSENBLATT (1961), *Distribution Free Tests of Independence Based on the Sample Distribution Function*, The Annals of Mathematical Statistics, Vol. 32, pp. 485-498. 
- [22] C. SCHENSTED (1961), *Longest increasing and decreasing subsequences*, Canadian Journal of Mathematics, Vol. 13, pp. 179-191. 
- [23] R. C. ELANDT (1962), *Exact and Approximate Power Function of the Non-Parametric Test of Tendency*, The Annals of Mathematical Statistics, Vol. 33, pp. 471-481. 
- [24] S. E. FELDMAN, E. KLINGER (1963), *Short cut calculation of the Fisher-Yates „exact test“*, Psychometrika, Vol. 28, pp. 289-291. 
- [25] H. D. SHANE, M. L. PURI (1969), *Rank Order Tests for Multivariate Paired Comparisons*, The Annals of Mathematical Statistics, Vol. 40, 2101-2117. 
- [26] M. L. PURI, P. K. SEN AND D. V. GOKHALE (1970), *On a Class of Rank Order Tests for Independence in Multivariate Distributions*, Sankhya: The Indian Journal of Statistics, Series A, Vol. 32, pp. 271-298. 

Bibliographie

- [27] B. K. SINHA, H. S. WIEAND (1977), *Multivariate Nonparametric Tests for Independence*, Journal Of Multivariate Analysis, Vol. 7, pp. 572-583. ☐
- [28] W. J. CONOVER, R. L. IMAN (1981), *Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics*, The American Statistician, Vol. 35, pp. 124-129. ☐
- [29] P. DEHEUVELS (1981), *An Asymptotic Decomposition for Multivariate Distribution-Free Tests of Independence*, Journal of Multivariate Analysis, Vol. 11, pp. 102-113. ☐
- [30] R. L. PLACKETT (1983), *Karl Pearson and the Chi-Squared Test*, International Statistical Review / Revue Internationale de Statistique, Vol. 51, pp. 59-72. ☐
- [31] S. CSÖRGÖ (1985), *Testing for Independence by the Empirical Characteristic Function*, Journal Of Multivariate Analysis, Vol. 16, pp. 290-299. ☐
- [32] J. L. RODGERS, W. A. NICEWANDER (1988), *Thirteen Ways to Look at the Correlation Coefficient*, The American Statistician, Vol. 42, pp. 59-66. ☐
- [33] M. H. DEGROOT (1988), *A Conversation with George A. Barnard*, Statistical Science, Vol. 3, pp. 196-212. ☐
- [34] R. H. RANGLES (1989), *A Distribution-Free Multivariate Sign Test Based on Interdirections*, Journal of the American Statistical Association, Vol. 84, pp. 1045-1050. ☐
- [35] A. FEUERVERGER (1993), *A Consistent Test for Bivariate Dependence*, International Statistical Review / Revue Internationale de Statistique, Vol. 61, pp. 419-433. ☐
- [36] A. ZOGHBI, I. STOJMENOVIC (1998), *Fast algorithms for generating integer partitions*, International Journal of Computer Mathematics, Vol. 70, pp 319-332. ☐
- [37] J BAIK, P. DEIFT AND K. JOHANSSON (1999), *On the distribution of the length of the longest increasing subsequence of random permutations*, Journal of the American Mathematical Society, Vol. 12, pp. 1119-1178. ☐
- [38] D. MARGARITIS, S. THRUN (2001), *A Bayesian Multiresolution Independence Test for Continuous Variables*, Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI), pp. 346-353, Morgan Kaufmann Publishers Inc., San Francisco. ☐
- [39] C. BANDT, B. POMPE (2002), *Permutation Entropy: A Natural Complexity Measure for Time Series*, Physical Review Letters, Vol. 88, pp. 174102. ☐
- [40] R. EDER, S. WENDELIN (2002), *Phenolzusammensetzung und antioxidative Kapazität von Trauben und Weinen*, Arbeitsgemeinschaft landwirtschaftlicher Versuchsanstalten, Jahrestagung 2002 in Klosterneuburg. ☐

- [41] N. K. BAKIROV, M. L. RIZZO, G. J. SZÉKELY (2006), *A multivariate nonparametric test of independence*, Journal of Multivariate Analysis, Vol. 97, pp. 1742-1756. 
- [42] A. BEJAN (2006), *Largest eigenvalues and sample covariance matrices*, Masterarbeit, University of Warwick. 
- [43] G. E. WILDING, G. S. MUDHOLKAR (2008), *Empirical approximations for Hoeffding's test of bivariate independence using two Weibull extensions*, Statistical Methodology, Vol. 5, pp. 160-170. 
- [44] M. MATILLA-GARCÍA, M. R. MARÍN (2008), *A non-parametric independence test using permutation entropy*, Journal of Econometrics, Vol. 144, pp. 139-155. 
- [45] J. E. GARCÍA, V. A. GONZÁLEZ-LÓPEZ (2009), *A Nonparametric Independence Test using Random Permutations*, Preprint, arXiv:0908.2794v2. 
- [46] A. NAZAROV, N. STEPANOVA (2009), *On asymptotic efficiency of multivariate version of Spearman's rho*, Preprint, arXiv:0906.1059v1. 
- [47] J.-F. QUESSY (2009), *Theoretical efficiency comparisons of independence tests based on multivariate versions of Spearman's rho*, Metrika, Vol. 70, pp. 315-338. 
- [48] K. M. AL ZOUKRA (2010), *Geometrie zufälliger Permutationen und statistische Anwendungen*, Bachelorarbeit, Greifswald.

Webdokumente

- [49] GERARD MICHON (NUMERICA.COM) (2000), *Partition Function*, <http://www.numericana.com/data/partition.htm>.
- [50] R. P. STANLEY (2008), *Hook Lengths and Contents*, <http://www-math.mit.edu/~rstan/transparencies/hooks.pdf>.
- [51] G. CARDILLO (2009), *MyBarnard: a very compact routine for Barnard's exact test on 2x2 matrix*, <http://www.mathworks.com/matlabcentral/fileexchange/25760>.
- [52] G. CARDILLO (2009), *Barnard's unconditional exact test on 2x2 matrix*, <http://www.advancedmcode.org/barnard.html>.
- [53] STEFAN AEGERHARD / UCI, *Wine recognition data*, <http://archive.ics.uci.edu/ml/datasets/Wine>.
- [54] WIKIPEDIA, *Phenole im Wein*, http://de.wikipedia.org/wiki/Phenole_im_Wein.

5 Anhang

5.1 Fliegendatensatz

Fliegendatensatz aus dem Datenarchiv des Instituts für Statistik und des SFB 386 der LMU München: <http://www.statistik.lmu.de/service/datenarchiv/fliegen/fliegen.html>

- Laenge** Körperlänge der Fliegen in mm
- Breite** Körperbreite der Fliegen in mm
- Fluegel** Flügellänge der Fliegen in mm
- E** Geweihlänge der Fliegen in mm
- WM** Geschlecht der Fliegen (0=weiblich,1=männlich)

Tabelle 5.1: Fliegendatensatz

Laenge	Breite	Fluegel	E	WM
3.30	NA	2.45	1.45	0
3.71	0.93	2.58	1.75	0
3.75	0.95	2.53	1.50	0
4.05	NA	NA	2	0
4.14	1.10	2.77	2.00	0
4.25	0.97	3	1.94	0
4.45	NA	2.71	2.05	0
4.60	1.15	3.1	2.20	0
4.72	1.22	3.1	2.16	0
5.00	1.19	2.95	2.25	0
5.11	1.43	3.6	2.63	0
5.20	1.42	3.32	2.60	0
3.31	1.00	2.38	1.75	1
3.60	1.00	2.45	1.95	1
3.98	1.09	2.68	2.21	1
4.08	1.14	2.88	2.52	1
4.30	1.25	3.2	3.25	1
4.35	1.25	3.08	2.80	1
4.40	NA	NA	2.50	1
4.40	NA	3.04	2.70	1
4.70	1.24	2.95	2.58	1

5.2 p-Werte des Korrelationsbeispiels

Die durchschnittlichen p-Werte der 21 Beispiele zur Korrelation von Seite 88:

No.	Pearson	Spearman	Chi2	Fisher	Kendall	Quadrant	Barnard	Hoefding	Bakrov	GarciaMarin	LIS	GRaFu	GRaPl
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0000	0.0000	0.0010	0.0859	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0002	0.0002	0.0000	0.0150	0.0001	0.0000	0.0010	0.2799	0.0009	0.0078	0.0172
3	0.0002	0.0008	0.0253	0.0214	0.0010	0.1374	0.0177	0.0028	0.0050	0.4712	0.1188	0.2663	0.2910
4	0.5181	0.5203	0.4813	0.3249	0.5272	0.6627	0.2915	0.4968	0.4856	0.4548	0.6915	0.5007	0.5123
5	0.0002	0.0008	0.0253	0.0214	0.0010	0.1374	0.0177	0.0028	0.0050	0.4872	0.1799	0.2663	0.2910
6	0.0000	0.0000	0.0002	0.0002	0.0000	0.0150	0.0001	0.0000	0.0010	0.3091	0.0096	0.0078	0.0172
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0015	0.0000	0.0000	0.0010	0.0927	0.0002	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000	0.0000
11	NaN	NaN	NaN	1.0000	1.0000	0.0000	0.9975	0.0000	0.0110	0.0010	0.0000	0.0000	0.0000
12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000	0.0000
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000	0.0000
14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0010	0.0010	0.0000	0.0000	0.0000
15	0.4896	0.4152	0.0871	0.0652	0.3917	0.2444	0.0564	0.0053	0.0010	0.2815	0.2500	0.0000	0.0001
16	0.6118	0.5422	0.4498	0.3069	0.5883	0.6448	0.2758	0.2948	0.1806	0.4666	0.5406	0.2490	0.2690
17	0.5695	0.5480	0.5037	0.3396	0.5639	0.6833	0.3049	0.3173	0.2976	0.5272	0.5894	0.3076	0.3227
18	0.4115	0.3892	0.3846	0.2640	0.3774	0.5763	0.2368	0.0048	0.0013	0.5236	0.4819	0.0574	0.0873
19	0.3550	0.3956	0.4839	0.3271	0.3606	0.6688	0.2934	0.0884	0.0036	0.4297	0.0781	0.0083	0.0211
20	0.6060	0.6428	0.5130	0.3422	0.7448	0.6808	0.3056	0.0424	0.0171	0.4543	0.6767	0.0010	0.0076
21	0.4886	0.5227	0.4709	0.3209	0.5332	0.6650	0.2886	0.5216	0.5561	0.4818	0.7209	0.4966	0.4828

5.3 Quantile der Verteilung von Hoeffdings D

In der Tabelle stehen die Quantilswerte für $30D$.

n	Statistische Sicherheit $1 - \alpha$					
	0.9	0.95	0.975	0.99	0.995	0.999
5	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0.1667	0.5000	0.5000	1.0000	1.0000	1.0000
7	0.1429	0.2857	0.4286	0.5714	0.7143	1.0000
8	0.1339	0.2321	0.3214	0.4286	0.5000	0.7143
9	0.1190	0.1905	0.2659	0.3651	0.4444	0.6032
10	0.1000	0.1587	0.2204	0.3047	0.3701	0.5239
11	0.0869	0.1374	0.1900	0.2635	0.3181	0.4446
12	0.0764	0.1206	0.1670	0.2323	0.2816	0.3989
13	0.0684	0.1078	0.1495	0.2062	0.2502	0.3514
14	0.0618	0.0970	0.1348	0.1865	0.2274	0.3187
15	0.0564	0.0890	0.1236	0.1713	0.2074	0.2936
16	0.0513	0.0808	0.1123	0.1557	0.1891	0.2673
17	0.0475	0.0750	0.1038	0.1443	0.1757	0.2473
18	0.0442	0.0699	0.0967	0.1341	0.1632	0.2292
19	0.0412	0.0651	0.0905	0.1248	0.1516	0.2137
20	0.0386	0.0608	0.0843	0.1171	0.1419	0.2015
21	0.0361	0.0571	0.0792	0.1092	0.1331	0.1895
22	0.0343	0.0540	0.0747	0.1034	0.1253	0.1773
23	0.0324	0.0511	0.0709	0.0981	0.1192	0.1684
24	0.0307	0.0484	0.0672	0.0929	0.1128	0.1617
25	0.0292	0.0462	0.0643	0.0888	0.1075	0.1534
26	0.0278	0.0440	0.0610	0.0845	0.1025	0.1458
27	0.0267	0.0422	0.0585	0.0811	0.0982	0.1392
28	0.0254	0.0401	0.0555	0.0773	0.0938	0.1336
29	0.0243	0.0385	0.0534	0.0739	0.0896	0.1273
30	0.0235	0.0371	0.0513	0.0710	0.0863	0.1228
31	0.0226	0.0357	0.0493	0.0686	0.0838	0.1186
32	0.0217	0.0343	0.0475	0.0656	0.0798	0.1139
33	0.0210	0.0331	0.0460	0.0639	0.0775	0.1106
34	0.0202	0.0319	0.0443	0.0612	0.0745	0.1060
35	0.0195	0.0308	0.0429	0.0593	0.0721	0.1018
40	0.0168	0.0266	0.0369	0.0510	0.0620	0.0887
45	0.0147	0.0232	0.0323	0.0448	0.0546	0.0772
50	0.0130	0.0206	0.0287	0.0396	0.0483	0.0689
60	0.0107	0.0169	0.0234	0.0324	0.0395	0.0558
70	0.0090	0.0143	0.0198	0.0274	0.0333	0.0476
80	0.0077	0.0124	0.0172	0.0239	0.0290	0.0415
90	0.0069	0.0109	0.0152	0.0211	0.0257	0.0361
100	0.0061	0.0097	0.0135	0.0188	0.0229	0.0326

5.4 Quantile der Verteilung von Bakirovs I_n

n	Statistische Sicherheit $1 - \alpha$					
	0.9	0.95	0.975	0.99	0.995	0.999
1	NaN	NaN	NaN	NaN	NaN	NaN
2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	0.7549	0.7549	0.7549	0.7549	0.7549	0.7549
4	0.6280	0.6678	0.6678	0.6678	0.6678	0.6678
5	0.5596	0.5911	0.6022	0.6256	0.6256	0.6256
6	0.5121	0.5398	0.5519	0.5717	0.5867	0.6019
7	0.4786	0.4976	0.5185	0.5422	0.5526	0.5766
8	0.4411	0.4682	0.4853	0.5085	0.5241	0.5480
9	0.4117	0.4390	0.4623	0.4845	0.4966	0.5242
10	0.3898	0.4133	0.4411	0.4622	0.4769	0.5058
11	0.3719	0.3945	0.4180	0.4433	0.4583	0.4848
12	0.3554	0.3782	0.4019	0.4292	0.4444	0.4724
13	0.3394	0.3622	0.3838	0.4093	0.4269	0.4554
14	0.3257	0.3483	0.3688	0.3948	0.4133	0.4434
15	0.3161	0.3372	0.3586	0.3812	0.3956	0.4237
16	0.3065	0.3277	0.3465	0.3683	0.3838	0.4155
17	0.2956	0.3165	0.3358	0.3611	0.3740	0.4033
18	0.2863	0.3067	0.3255	0.3507	0.3678	0.4010
19	0.2784	0.2976	0.3179	0.3414	0.3589	0.3893
20	0.2724	0.2919	0.3095	0.3315	0.3449	0.3778
21	0.2660	0.2847	0.3018	0.3234	0.3404	0.3713
22	0.2595	0.2778	0.2954	0.3173	0.3336	0.3627
23	0.2535	0.2711	0.2869	0.3083	0.3264	0.3581
24	0.2490	0.2663	0.2825	0.3030	0.3176	0.3529
25	0.2439	0.2612	0.2782	0.3010	0.3142	0.3443
26	0.2381	0.2561	0.2725	0.2919	0.3090	0.3361
27	0.2337	0.2514	0.2677	0.2872	0.2990	0.3334
28	0.2299	0.2473	0.2631	0.2837	0.3004	0.3295
29	0.2252	0.2411	0.2570	0.2764	0.2893	0.3207
30	0.2221	0.2393	0.2555	0.2736	0.2883	0.3194
35	0.2050	0.2196	0.2340	0.2518	0.2628	0.2894
40	0.1920	0.2068	0.2199	0.2355	0.2466	0.2728
45	0.1805	0.1938	0.2074	0.2237	0.2389	0.2644
50	0.1720	0.1862	0.1984	0.2134	0.2241	0.2486
60	0.1564	0.1686	0.1797	0.1934	0.2059	0.2237
70	0.1448	0.1571	0.1675	0.1817	0.1902	0.2102
80	0.1358	0.1461	0.1567	0.1692	0.1777	0.1975
90	0.1275	0.1376	0.1475	0.1609	0.1699	0.1883
100	0.1216	0.1307	0.1392	0.1502	0.1577	0.1727

5.5 Quantile der Teststatistik G des Permutation-Entropy-Tests

Es folgen die Quantiltabellen der Teststatistik G für verschiedene Stichprobenumfänge n für eine Bindungsdimension $m=3$ und $m=4$. Ist darauf hinzuweisen, dass die Quantile nicht die gewünschte Genauigkeit besitzen, da die Verteilungen für Zufallspermutationen ziemlich stark variieren. Die Quantile sind daher mit Vorsicht zu behandeln. Zur Sicherheit wird daher im Matlabprogramm [2.3.6](#) ein Konfidenzintervall auf Basis der Varianz gebildet.

m=3 n	Statistische Sicherheit $1 - \alpha$			
	0.9	0.95	0.975	0.99
6	6.02	8.30	9.63	9.84
7	7.37	8.41	8.41	9.39
8	8.53	9.36	9.41	11.09
9	7.68	9.98	10.51	11.64
10	8.37	10.25	11.40	12.03
11	8.77	9.94	12.34	13.23
12	9.01	10.24	11.64	14.09
13	8.66	10.83	11.77	14.81
14	8.28	10.85	12.26	13.92
15	8.44	10.33	13.04	14.28
16	8.53	10.03	12.54	14.68
17	9.00	10.60	12.34	15.42
18	8.71	10.35	12.08	15.03
19	8.53	10.59	12.35	15.21
20	8.29	10.51	12.10	14.90
21	8.25	10.44	12.46	15.08
22	8.25	10.25	12.41	14.83
23	7.96	10.26	12.47	15.19
24	7.70	10.11	11.98	14.98
25	7.52	9.96	12.16	14.85
26	7.55	9.98	12.29	15.02
27	7.56	9.86	12.03	14.88
28	7.51	9.63	12.08	14.74
29	7.43	9.57	12.13	15.06
30	7.57	9.72	12.05	15.26
35	7.43	9.50	11.60	14.81
40	7.34	9.45	11.48	14.70
45	7.33	9.36	11.41	14.07
50	7.19	9.16	11.14	14.15
60	7.19	9.21	11.37	14.14
70	7.16	9.13	10.98	13.75
80	7.13	9.22	11.12	13.89
90	7.13	9.14	11.12	13.85
100	7.16	9.03	11.09	13.84

5.5 Quantile der Teststatistik G des Permutation-Entropy-Tests

m=4 n	Statistische Sicherheit $1 - \alpha$			
	0.9	0.95	0.975	0.99
24	32.36	34.86	37.33	40.13
25	32.32	35.04	37.49	40.33
26	32.64	35.26	37.66	40.73
27	32.94	35.63	38.38	41.78
28	33.10	36.13	38.64	42.20
29	33.16	36.14	38.89	42.40
30	33.37	36.50	39.17	42.76
31	33.66	36.73	39.76	43.42
32	33.78	36.95	39.79	43.63
33	33.76	37.09	40.12	43.95
34	33.90	37.32	40.28	44.55
35	34.24	37.52	40.59	44.55
36	34.31	37.60	40.57	44.44
37	34.13	37.59	40.72	44.62
38	34.33	37.53	40.97	44.75
39	34.41	37.90	40.97	44.54
40	34.45	37.91	41.29	45.29
41	34.45	38.20	41.47	45.42
42	34.40	37.75	41.16	45.39
43	34.42	38.06	41.42	45.65
44	34.36	38.12	41.36	45.15
45	34.56	38.30	41.65	46.00
46	34.23	38.00	41.73	45.73
47	34.56	38.45	42.10	46.10
48	34.26	37.95	41.36	46.04
49	34.52	38.23	41.79	45.82
50	34.33	38.12	41.39	45.83
60	34.07	37.93	41.64	46.46
70	34.17	38.19	42.15	46.53
80	33.43	37.62	41.47	46.34
90	33.03	37.25	41.01	46.29
100	32.55	36.88	40.96	46.02

5.6 Quantile der Länge der längsten aufsteigenden Teilfolge L_n (zweiseitig)

n	Quantile							
	0.005	0.0125	0.025	0.05	0.95	0.975	0.9875	0.995
1	1	1	1	1	1	1	1	1
2	1	1	1	1	2	2	2	2
3	1	1	1	1	3	3	3	3
4	1	1	1	2	3	4	4	4
5	1	2	2	2	4	4	4	5
6	2	2	2	2	4	5	5	5
7	2	2	2	2	5	5	5	6
8	2	2	2	3	5	5	6	6
9	2	2	3	3	5	6	6	6
10	3	3	3	3	6	6	6	7
11	3	3	3	3	6	6	7	7
12	3	3	3	3	6	7	7	8
13	3	3	3	4	7	7	7	8
14	3	3	4	4	7	7	8	8
15	3	4	4	4	7	8	8	8
16	4	4	4	4	8	8	8	9
17	4	4	4	4	8	8	9	9
18	4	4	4	5	8	8	9	9
19	4	4	5	5	8	9	9	9
20	4	5	5	5	8	9	9	10
21	4	5	5	5	9	9	9	10
22	5	5	5	5	9	9	10	10
23	5	5	5	6	9	10	10	10
24	5	5	5	6	9	10	10	11
25	5	5	6	6	10	10	10	11
26	5	5	6	6	10	10	11	11
27	5	6	6	6	10	10	11	11
28	6	6	6	6	10	11	11	11
29	6	6	6	6	10	11	11	12
30	6	6	6	7	10	11	11	12
35	6	7	7	7	11	12	12	13
40	7	7	8	8	12	13	13	14
45	8	8	8	9	13	13	14	15
50	8	9	9	9	14	14	15	15
60	9	10	10	10	15	16	16	17
70	10	11	11	11	16	17	17	18
80	11	12	12	12	17	18	19	19
90	12	13	13	13	19	19	20	20
100	13	13	14	14	20	20	21	21

5.7 Quantile von S_u und S_f des Geometrietests

Es folgen die Quantilstabellen für die Anzahlen S_u und S_f des Geometrietests für $k=1, 2, 3$ und 4. Für $n=1, \dots, 9$ wurden dabei alle möglichen Permutationen untersucht. Für $n \geq 10$ entstanden die Quantile aus der 50-maligen Simulation von 10000 Zufallspermutationen und sind aus dem arithmetischen Mittel der 50 Simulationsquantilen geschätzt. Da für $k=1$ die Varianz der Quantile für S_u im Durchschnitt bei 0.0509 liegt, weichen die wahren Quantile mit 95%-iger Sicherheit um maximal ± 0.0625 von den gelisteten ab. Ähnlich kleine Abweichungen erhalten wir auch für S_f und andere k .

k=1 n	Quantile von S_u					
	0.9	0.95	0.975	0.99	0.995	0.999
6	16.00	16.00	16.00	16.00	16.00	16.00
7	24.00	24.00	25.00	25.00	25.00	25.00
8	32.00	33.00	34.00	34.00	35.00	36.00
9	41.00	42.00	43.00	44.00	44.00	45.00
10	49.62	50.99	51.91	52.98	53.37	54.98
11	58.00	59.24	60.57	61.97	62.60	64.20
12	67.00	68.00	69.19	70.97	71.80	73.34
13	75.95	77.00	78.02	79.91	80.76	82.53
14	84.05	86.00	87.00	88.80	89.66	91.50
15	93.00	95.00	96.00	97.59	98.66	100.63
16	102.00	103.87	105.00	106.51	107.53	109.52
17	111.00	112.30	114.00	115.22	116.36	118.43
18	119.99	121.03	122.97	124.11	125.24	127.52
19	128.75	130.00	131.88	133.06	134.15	136.38
20	137.10	139.00	140.56	142.01	143.07	145.27
21	146.00	148.00	149.44	151.00	152.08	154.29
22	155.00	157.00	158.19	160.00	161.06	163.26
23	164.00	166.00	167.03	169.00	170.00	172.16
24	173.00	174.99	176.00	178.00	179.00	181.15
25	182.00	183.92	185.00	186.97	187.97	190.07
26	191.00	192.94	194.00	195.96	196.97	198.95
27	200.00	201.66	203.00	204.93	205.93	207.96
28	208.99	210.41	212.00	213.84	214.86	216.91
29	217.93	219.25	221.00	222.76	223.90	225.91
30	226.68	228.09	230.00	231.63	232.74	234.72
35	271.01	273.00	274.94	276.23	277.31	279.65
40	316.00	318.00	319.52	321.01	322.10	324.41
45	361.00	363.00	364.23	366.00	367.07	369.26
50	406.00	407.99	409.05	411.00	412.01	414.18
60	495.96	497.72	499.00	500.96	501.96	503.96
70	585.68	587.35	589.00	590.83	591.93	593.87
80	675.11	677.03	679.00	680.62	681.79	683.72
90	765.06	767.02	768.99	770.35	771.67	773.57
100	855.02	857.02	858.95	860.26	861.45	863.56

5 Anhang

k=1 n	Quantile von S_f					
	0.9	0.95	0.975	0.99	0.995	0.999
6	16.00	17.00	17.00	17.00	18.00	18.00
7	20.00	21.00	21.00	21.00	21.00	22.00
8	24.00	25.00	25.00	25.00	25.00	26.00
9	28.00	29.00	29.00	29.00	29.00	30.00
10	32.00	32.69	33.00	33.00	33.00	34.00
11	36.00	36.08	37.00	37.00	37.00	37.99
12	40.00	40.01	41.00	41.00	41.00	41.93
13	44.00	44.00	45.00	45.00	45.00	45.90
14	48.00	48.00	49.00	49.00	49.00	49.76
15	52.00	52.00	53.00	53.00	53.00	53.64
16	56.00	56.00	57.00	57.00	57.00	57.42
17	60.00	60.00	61.00	61.00	61.00	61.32
18	64.00	64.00	65.00	65.00	65.00	65.19
19	68.00	68.00	68.96	69.00	69.00	69.17
20	72.00	72.00	72.81	73.00	73.00	73.11
21	76.00	76.00	76.56	77.00	77.00	77.06
22	80.00	80.00	80.31	81.00	81.00	81.04
23	84.00	84.00	84.08	85.00	85.00	85.04
24	88.00	88.00	88.00	89.00	89.00	89.02
25	92.00	92.00	92.00	93.00	93.00	93.00
26	96.00	96.00	96.01	97.00	97.00	97.00
27	100.00	100.00	100.00	101.00	101.00	101.00
28	104.00	104.00	104.00	105.00	105.00	105.01
29	108.00	108.00	108.00	109.00	109.00	109.00
30	112.00	112.00	112.00	113.00	113.00	113.00
35	132.00	132.00	132.00	133.00	133.00	133.00
40	152.00	152.00	152.00	153.00	153.00	153.00
45	172.00	172.00	172.00	172.97	173.00	173.00
50	192.00	192.00	192.00	192.78	193.00	193.00
60	232.00	232.00	232.00	232.13	233.00	233.00
70	272.00	272.00	272.00	272.01	273.00	273.00
80	312.00	312.00	312.00	312.00	312.96	313.00
90	352.00	352.00	352.00	352.00	352.88	353.00
100	392.00	392.00	392.00	392.00	392.69	393.00

5.7 Quantile von S_u und S_f des Geometrietests

k=2 n	Quantile von S_u					
	0.9	0.95	0.975	0.99	0.995	0.999
10	36.00	36.00	36.00	36.00	36.00	36.00
11	49.00	49.00	49.00	49.00	49.00	49.00
12	64.00	64.00	64.00	64.00	64.00	64.00
13	81.00	81.00	81.00	81.00	81.00	81.00
14	98.94	99.00	100.00	100.00	100.00	100.00
15	117.03	118.83	119.01	120.00	121.00	121.00
16	137.00	138.96	140.00	141.00	142.00	143.13
17	157.02	159.25	161.00	162.99	163.99	165.96
18	178.00	180.91	182.94	185.01	186.54	189.26
19	199.17	202.29	205.00	207.72	209.43	212.79
20	221.00	224.47	227.35	230.66	232.76	236.87
21	242.98	246.98	250.21	253.97	256.26	261.00
22	265.03	269.44	273.11	277.23	280.05	285.56
23	287.68	292.29	296.32	300.89	303.85	309.99
24	310.18	315.36	319.72	324.48	327.73	334.26
25	333.04	338.48	343.11	348.36	351.79	358.75
26	356.07	361.91	366.74	372.24	375.94	383.35
27	379.26	385.27	390.38	396.14	400.01	407.55
28	402.52	408.83	414.19	420.17	424.28	432.64
29	425.96	432.42	438.05	444.38	448.69	457.44
30	449.44	456.14	461.90	468.45	472.85	481.55
35	568.22	575.99	582.60	590.09	595.17	605.36
40	688.48	696.95	704.13	712.48	717.95	729.38
45	809.70	818.68	826.51	835.25	841.12	852.96
50	931.62	941.17	949.29	958.46	964.70	976.77
60	1177.02	1187.18	1195.85	1205.71	1212.10	1224.71
70	1423.84	1434.49	1443.45	1453.62	1460.36	1474.07
80	1671.18	1682.17	1691.48	1702.02	1709.08	1722.83
90	1919.39	1930.61	1940.12	1950.92	1958.13	1972.22
100	2167.72	2179.20	2188.82	2199.85	2207.11	2221.30

5 Anhang

k=2 n	Quantile von S_f					
	0.9	0.95	0.975	0.99	0.995	0.999
10	48.00	49.00	49.00	49.00	49.00	49.00
11	61.02	62.00	63.00	63.23	64.00	64.00
12	75.00	76.00	77.00	78.01	79.00	79.99
13	89.00	90.97	92.00	93.26	94.17	96.04
14	103.33	105.14	107.00	108.90	109.99	112.19
15	118.00	120.04	122.01	124.11	125.74	128.46
16	132.99	135.10	137.25	139.92	141.45	144.60
17	147.92	150.32	152.92	155.37	157.20	160.72
18	162.92	165.87	168.04	171.04	173.03	177.01
19	178.00	181.00	183.76	186.80	188.82	192.96
20	193.00	196.26	199.10	202.40	204.61	208.99
21	208.15	211.96	214.94	218.16	220.46	225.24
22	223.77	227.07	230.22	233.95	236.29	241.10
23	239.00	242.91	246.00	249.66	252.04	256.95
24	254.55	258.23	261.73	265.39	267.85	272.87
25	269.99	273.96	277.26	281.11	283.73	289.12
26	285.60	289.64	293.02	296.88	299.46	304.68
27	301.04	305.12	308.77	312.70	315.34	320.88
28	316.70	320.89	324.26	328.42	331.16	336.70
29	332.07	336.54	340.11	344.31	347.07	352.67
30	347.92	352.13	355.93	360.05	362.85	368.41
35	426.24	430.98	434.98	439.25	442.32	448.31
40	505.15	510.01	514.10	518.64	521.71	527.49
45	584.24	589.22	593.41	598.14	601.26	607.43
50	663.63	668.65	672.97	677.69	680.83	686.99
60	822.42	827.66	832.02	836.94	840.18	846.40
70	981.73	987.00	991.40	996.41	999.69	1006.07
80	1140.97	1146.41	1150.94	1156.03	1159.38	1165.79
90	1300.69	1306.04	1310.65	1315.70	1319.08	1325.74
100	1460.12	1465.74	1470.28	1475.36	1478.72	1485.26

5.7 Quantile von S_u und S_f des Geometrietests

k=3 n	Quantile von S_u					
	0.9	0.95	0.975	0.99	0.995	0.999
15	81.00	81.00	81.00	81.00	81.00	81.00
16	100.00	100.00	100.00	100.00	100.00	100.00
17	121.00	121.00	121.00	121.00	121.00	121.00
18	144.00	144.00	144.00	144.00	144.00	144.00
19	169.00	169.00	169.00	169.00	169.00	169.00
20	196.00	196.00	196.00	196.00	196.00	196.00
21	225.00	225.00	225.00	225.00	225.00	225.00
22	254.03	255.19	256.00	256.00	256.00	256.00
23	286.00	287.00	288.00	289.00	289.00	289.00
24	318.00	320.00	321.64	322.99	323.56	324.00
25	352.00	354.14	356.02	358.00	359.01	360.75
26	386.30	389.72	392.00	394.32	395.88	398.05
27	421.99	425.89	428.82	431.86	433.57	436.69
28	458.15	462.90	466.26	470.10	472.30	476.39
29	495.37	500.67	504.73	509.12	511.87	516.82
30	533.16	539.13	543.95	549.03	552.29	558.31
35	730.92	740.21	747.93	756.70	762.33	773.25
40	939.35	951.67	962.07	973.78	981.62	997.61
45	1155.25	1170.35	1183.11	1197.55	1207.24	1227.43
50	1376.40	1393.79	1408.67	1425.46	1437.05	1459.86
60	1830.17	1851.14	1869.12	1889.47	1903.43	1931.90
70	2293.68	2317.65	2338.11	2361.45	2377.17	2409.61
80	2763.14	2789.03	2811.36	2836.78	2853.80	2888.90
90	3237.42	3265.24	3288.68	3315.98	3334.20	3371.30
100	3714.65	3743.72	3768.61	3797.07	3816.15	3855.49

k=3 n	Quantile von S_f					
	0.9	0.95	0.975	0.99	0.995	0.999
15	100.00	100.00	100.00	100.00	100.00	100.00
16	121.00	121.00	121.00	121.00	121.00	121.00
17	144.00	144.00	144.00	144.00	144.00	144.00
18	167.00	168.00	169.00	169.00	169.00	169.00
19	192.00	193.99	194.78	195.50	196.00	196.00
20	218.00	220.00	221.52	223.00	223.76	224.99
21	245.00	247.55	249.23	251.15	252.29	254.21
22	272.88	275.94	278.02	280.70	282.07	284.79
23	301.00	304.77	307.68	310.70	312.56	316.07
24	329.95	334.10	337.57	341.19	343.55	348.01
25	359.37	364.21	368.12	372.46	375.15	380.39
26	389.31	394.80	399.13	404.02	407.21	413.19
27	419.77	425.64	430.53	435.99	439.51	446.45
28	450.37	456.90	462.25	468.29	472.29	480.06
29	481.41	488.51	494.28	500.89	505.15	513.43
30	512.92	520.36	526.66	533.74	538.43	547.76
35	673.59	683.32	691.49	701.08	707.21	719.89
40	838.82	850.20	860.01	871.17	878.65	893.80
45	1007.22	1020.08	1031.18	1043.73	1052.12	1069.23
50	1177.78	1191.80	1203.91	1217.71	1227.00	1245.42
60	1523.53	1539.33	1552.90	1568.24	1578.69	1600.29
70	1873.41	1890.60	1905.23	1921.87	1932.98	1955.69
80	2225.12	2243.40	2258.74	2276.61	2288.49	2312.34
90	2579.23	2598.12	2614.20	2632.61	2644.82	2669.59
100	2934.29	2953.76	2970.45	2989.53	3002.05	3028.05

5.7 Quantile von S_u und S_f des Geometrietests

k=4 n	Quantile von S_u					
	0.9	0.95	0.975	0.99	0.995	0.999
20	144.00	144.00	144.00	144.00	144.00	144.00
21	169.00	169.00	169.00	169.00	169.00	169.00
22	196.00	196.00	196.00	196.00	196.00	196.00
23	225.00	225.00	225.00	225.00	225.00	225.00
24	256.00	256.00	256.00	256.00	256.00	256.00
25	289.00	289.00	289.00	289.00	289.00	289.00
26	324.00	324.00	324.00	324.00	324.00	324.00
27	361.00	361.00	361.00	361.00	361.00	361.00
28	400.00	400.00	400.00	400.00	400.00	400.00
29	441.00	441.00	441.00	441.00	441.00	441.00
30	483.73	484.00	484.00	484.00	484.00	484.00
35	718.28	721.91	724.01	726.11	727.34	729.00
40	984.80	992.03	997.74	1003.43	1007.03	1013.03
45	1275.23	1287.01	1296.35	1306.49	1312.83	1324.84
50	1584.37	1600.46	1613.89	1628.43	1637.96	1656.49
60	2242.91	2267.62	2288.34	2311.80	2327.47	2358.16
70	2939.61	2971.51	2998.53	3029.63	3050.42	3091.88
80	3660.60	3699.03	3731.55	3769.07	3794.49	3846.59
90	4400.95	4444.31	4481.76	4524.48	4553.69	4610.41
100	5152.57	5200.40	5241.51	5288.52	5319.98	5385.02

k=4 n	Quantile von S_f					
	0.9	0.95	0.975	0.99	0.995	0.999
20	169.00	169.00	169.00	169.00	169.00	169.00
21	196.00	196.00	196.00	196.00	196.00	196.00
22	225.00	225.00	225.00	225.00	225.00	225.00
23	256.00	256.00	256.00	256.00	256.00	256.00
24	289.00	289.00	289.00	289.00	289.00	289.00
25	324.00	324.00	324.00	324.00	324.00	324.00
26	360.00	361.00	361.00	361.00	361.00	361.00
27	397.76	399.00	400.00	400.00	400.00	400.00
28	436.82	438.27	439.89	440.99	441.00	441.00
29	477.00	479.23	481.00	482.68	483.32	484.00
30	518.51	521.63	523.94	525.96	527.00	528.85
35	742.01	748.87	754.11	759.64	763.11	769.32
40	985.60	996.34	1005.14	1014.79	1021.03	1033.02
45	1243.78	1258.45	1270.58	1284.18	1292.98	1310.59
50	1512.88	1531.05	1546.22	1563.53	1575.18	1598.16
60	2073.70	2098.14	2118.72	2142.27	2158.05	2189.61
70	2655.70	2684.76	2709.74	2738.14	2756.82	2796.76
80	3250.42	3283.66	3311.77	3344.24	3366.30	3411.14
90	3855.40	3891.82	3922.86	3958.91	3983.45	4031.01
100	4466.21	4505.19	4538.53	4576.49	4602.13	4654.64