

GRaP Independence Test

Marcus Vollmer

Institut für Mathematik und Informatik
Universität Greifswald

17. Nachwuchsworkshop der DStatG
Leipzig

19.09.2011



Content

- 1 Problemstellung
- 2 GRaP Independence Test
- 3 Berechnung des p-Wertes
- 4 Power Analyse
 - Unabhängigkeit
 - Abhängigkeit
- 5 Beispiel
- 6 Zusammenfassung
- 7 Referenzen
- 8 Appendix



Problemstellung



Problemstellung

- stetige Merkmale X , Y

z.B. Tanne:

X - Höhe des Baumes

Y - Länge der Baumkrone



Problemstellung

- stetige Merkmale X , Y

z.B. Tanne: X - Höhe des Baumes
 Y - Länge der Baumkrone

- Testen die Hypothesen:

H_0 : X und Y sind unabhängig

H_1 : X und Y sind nicht unabhängig



Problemstellung

- stetige Merkmale X , Y

z.B. Tanne: X - Höhe des Baumes
 Y - Länge der Baumkrone

- Testen die Hypothesen:

H_0 : X und Y sind unabhängig

H_1 : X und Y sind nicht unabhängig

- empirische Ausprägungen x_1, \dots, x_n und y_1, \dots, y_n



Ein Blick auf die Unabhängigkeitstests

1895	Pearsons r
1904	Spearman's ρ
1938	Kendalls τ
1922	χ^2 -Test
1925	Fisher-Yates-Test
1945	Barnards CSM-Test
1962	Quadrantentest
1948	Hoeffdings D-Test
1993	Feuerverger-Test
2004	Bakirovs I_n
2008	Entropy-Test
2009	LIS-Test
2010	GRaP-Test



Ein Blick auf die Unabhängigkeitstests

1895	Pearsons r	}	Lineare / Monotone Abhängigkeit
1904	Spearman's ρ		
1938	Kendalls τ		
1922	χ^2 -Test		
1925	Fisher-Yates-Test		
1945	Barnards CSM-Test		
1962	Quadrantentest		
1948	Hoeffdings D-Test		
1993	Feuerverger-Test		
2004	Bakirovs I_n		
2008	Entropy-Test		
2009	LIS-Test		
2010	GRaP-Test		



Ein Blick auf die Unabhängigkeitstests

1895	Pearsons r	}	Lineare / Monotone Abhängigkeit
1904	Spearman's ρ		
1938	Kendalls τ		
1922	χ^2 -Test	}	Kategoriell / Häufigkeiten
1925	Fisher-Yates-Test		
1945	Barnards CSM-Test		
1962	Quadrantentest		
1948	Hoeffdings D-Test		
1993	Feuerverger-Test		
2004	Bakirovs I_n		
2008	Entropy-Test		
2009	LIS-Test		
2010	GRaP-Test		



Ein Blick auf die Unabhängigkeitstests

1895	Pearsons r	}	Lineare / Monotone Abhängigkeit
1904	Spearman's ρ		
1938	Kendalls τ		
1922	χ^2 -Test	}	Kategoriell / Häufigkeiten
1925	Fisher-Yates-Test		
1945	Barnards CSM-Test		
1962	Quadrantentest		
1948	Hoeffdings D-Test	}	Ordinal / Verschiedene Abhängigkeiten
1993	Feuerverger-Test		
2004	Bakirovs I_n		
2008	Entropy-Test		
2009	LIS-Test		
2010	GRaP-Test		



Ein Blick auf die Unabhängigkeitstests

1895	Pearsons r	}	Lineare / Monotone Abhängigkeit
1904	Spearman's ρ		
1938	Kendalls τ		
1922	χ^2 -Test	}	Kategoriell / Häufigkeiten
1925	Fisher-Yates-Test		
1945	Barnards CSM-Test		
1962	Quadrantentest		
1948	Hoeffdings D-Test	}	Unabhängigkeit stetiger Größen als ordinales Problem, denn $\phi : \mathbb{R} \rightarrow \mathbb{R}$ monoton, so gilt: X, Y unabh. $\Leftrightarrow \phi(X), Y$ unabh.
1993	Feuerverger-Test		
2004	Bakirovs I_n		
2008	Entropy-Test		
2009	LIS-Test		
2010	GRaP-Test		



Ein Blick auf die Unabhängigkeitstests

Hoeffding, Feuerverger, Bakirov

- Vergleich gemeinsame Verteilungsfunktion $F(x, y)$ mit Randverteilungen $F(x), F(y)$
- Testgröße: $\|F(x, y) - F(x)F(y)\|$



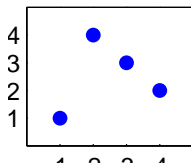
Ein Blick auf die Unabhängigkeitstests

Hoeffding, Feuerverger, Bakirov

- Vergleich gemeinsame Verteilungsfunktion $F(x, y)$ mit Randverteilungen $F(x), F(y)$
- Testgröße: $\|F(x, y) - F(x)F(y)\|$

Entropie, LIS, GRaP

- Zufällige Permutationen
- Streudiagramm der Rangzahlen ist eine Permutation



$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}$$

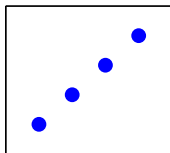
$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- Unabhängigkeit impliziert Gleichverteilung der Permutation



Testgröße für Permutationen?

z.B.



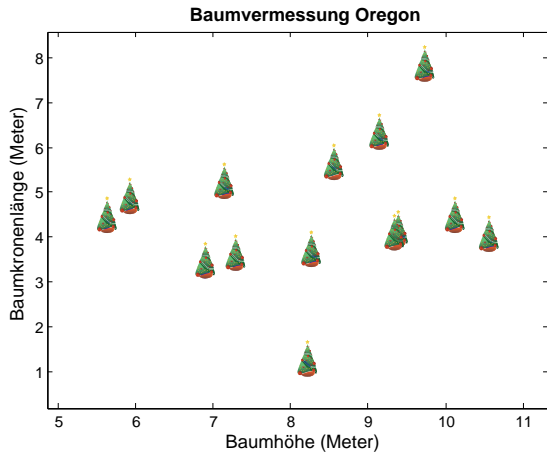
soll im Ablehnungsbereich liegen

- **Entropie-Test:** Zählen der ordinalen Muster (Zeitreihen)
- **LIS-Test:** maximale Länge aufsteigender Teilfolgen
- **GRaP:** "Geometry of Random Permutations"
 - "Fläche", die von Umgebungen der Punkte $(i, \pi(i))$ überdeckt wird
 - von Prof. Christoph Bandt und Marcus Vollmer
 - Für kleine Stichproben



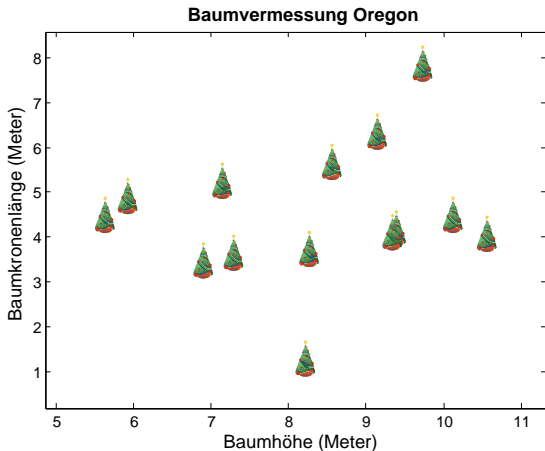
Nadelgehölze im Südwesten Oregons

$n=14$



Nadelgehölze im Südwesten Oregons

$n=14$



Pearsonkorrelation: $r = 0.19$
 Spearmankorrelation: $r_s = 0.14$
 Kendalls Tau: $\tau = 0.16$

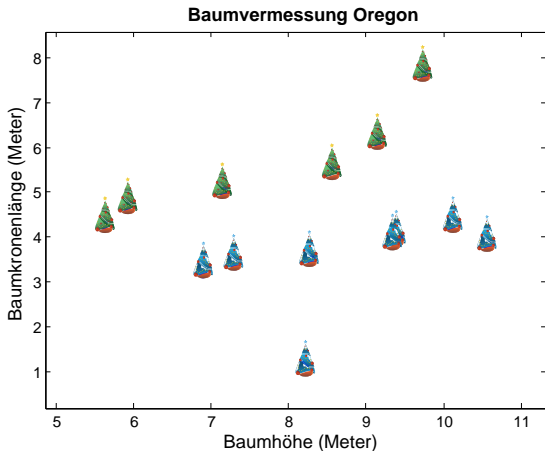
p-Werte:

Pearson = 0.516
 Spearman = 0.627
 Fisher = 0.500
 Kendall = 0.451
 Quadranten = 0.596
 Barnard = 0.395
 Hoeffding = 0.209
 Feuerverger = 0.460
 Bakirov = 0.555
 Entropie-Test = 0.968
 LIS = 0.807
 GRaP = 0.118



Nadelgehölze im Südwesten Oregons

$n=14$



Pearsonkorrelation: $r = 0.19$
 Spearmankorrelation: $r_s = 0.14$
 Kendalls Tau: $\tau = 0.16$

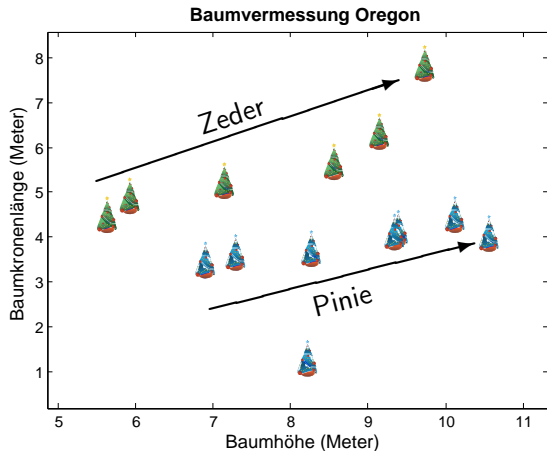
p-Werte:

Pearson = 0.516
 Spearman = 0.627
 Fisher = 0.500
 Kendall = 0.451
 Quadranten = 0.596
 Barnard = 0.395
 Hoeffding = 0.209
 Feuerverger = 0.460
 Bakirov = 0.555
 Entropie-Test = 0.968
 LIS = 0.807
 GRaP = 0.118



Nadelgehölze im Südwesten Oregons

$n=14$



Pearsonkorrelation: $r = 0.19$
 Spearmankorrelation: $r_s = 0.14$
 Kendalls Tau: $\tau = 0.16$

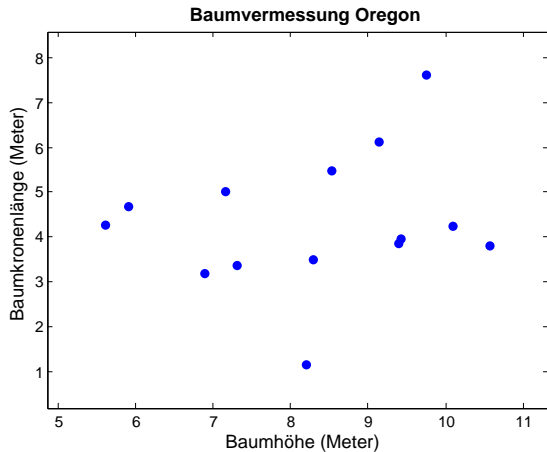
p-Werte:

Pearson = 0.516
 Spearman = 0.627
 Fisher = 0.500
 Kendall = 0.451
 Quadranten = 0.596
 Barnard = 0.395
 Hoeffding = 0.209
 Feuerverger = 0.460
 Bakirov = 0.555
 Entropie-Test = 0.968
 LIS = 0.807
 GRaP = 0.118



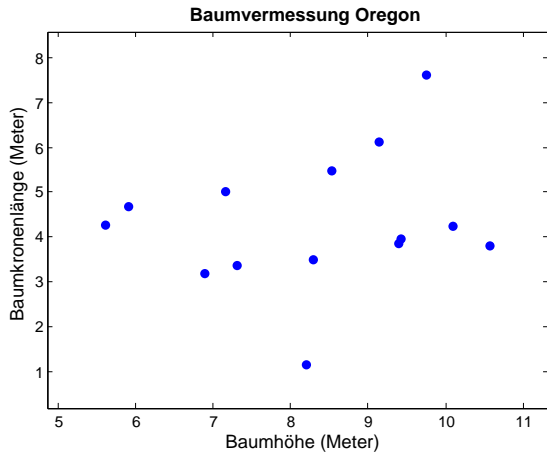
GRaP Independence Test

$n=14$



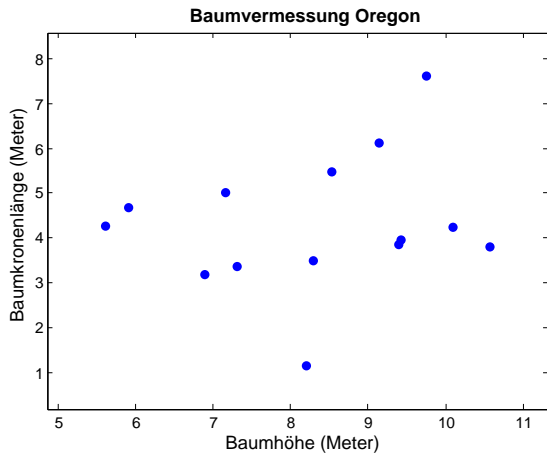
GRaP Independence Test

$n=14$



GRaP Independence Test

$n=14$

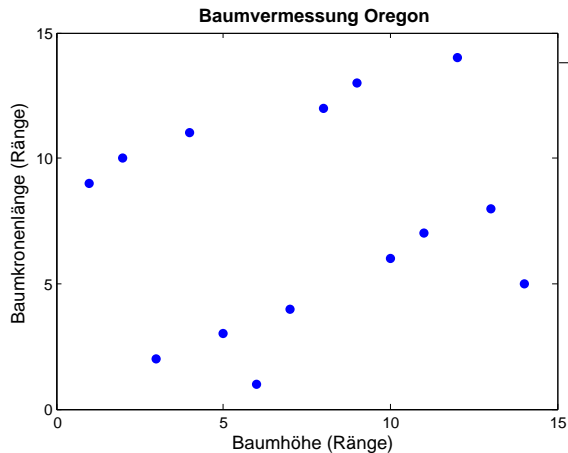


→ Rangtransformation



GRaP Independence Test

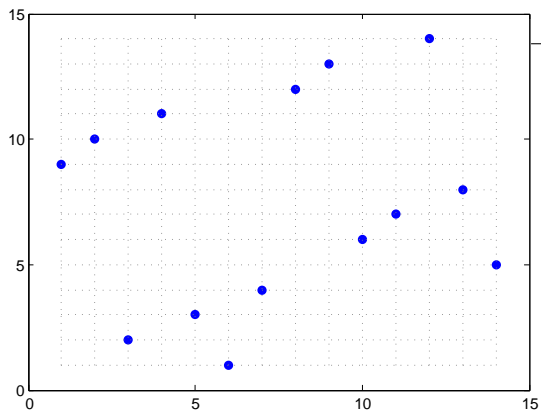
$n=14$



→ Rangtransformation



GRaP Independence Test

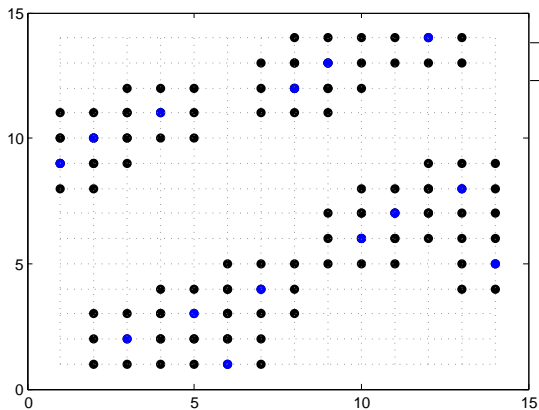
 $n=14$ 

Rangtransformation



GRaP Independence Test

$$n=14, r=1$$

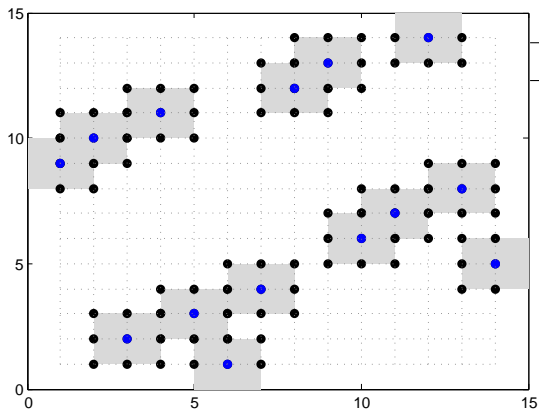


→ Rangtransformation
 → Zählen der
 überdeckten Punkte



GRaP Independence Test

$$n=14, r=1$$

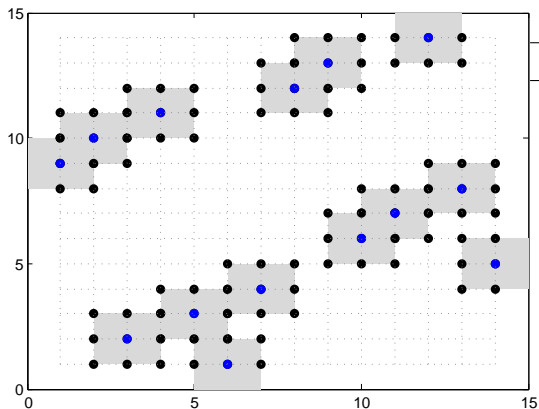


→ Rangtransformation
 → Zählen der
 überdeckten Punkte



GRaP Independence Test

$$n=14, r=1$$

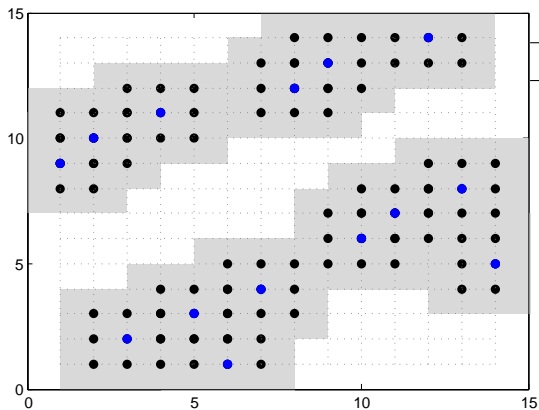


→ Rangtransformation
 → Zählen der
 überdeckten Punkte
 Anzahl: $s_1=92$



GRaP Independence Test

$$n=14, r=2$$

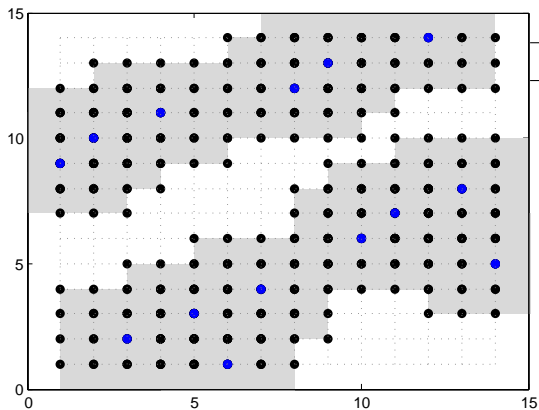


→ Rangtransformation
 → Zählen der überdeckten Punkte
 Anzahl: $s_1=92$



GRaP Independence Test

$$n=14, r=2$$

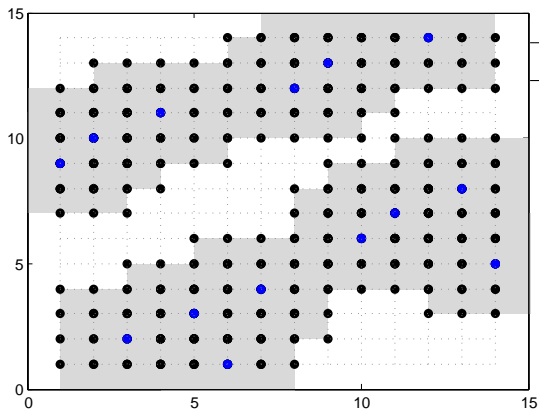


→ Rangtransformation
 → Zählen der
 überdeckten Punkte
 Anzahl: $s_1=92$



GRaP Independence Test

$$n=14, r=2$$

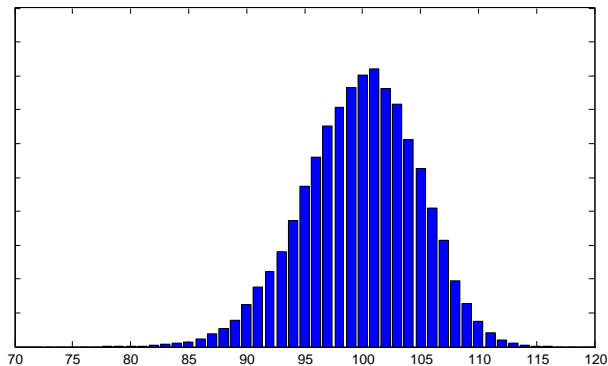


→ Rangtransformation
 → Zählen der
 überdeckten Punkte
 Anzahl: $s_1=92$
 $s_2=159$



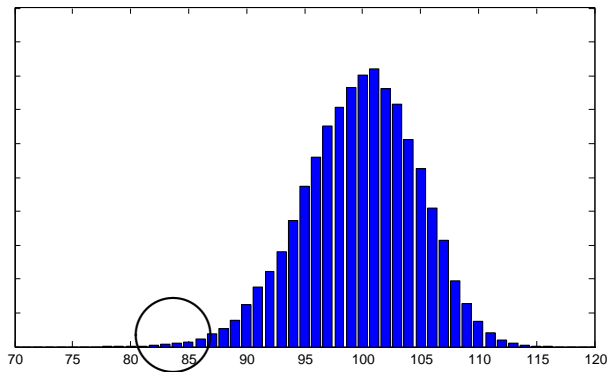
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



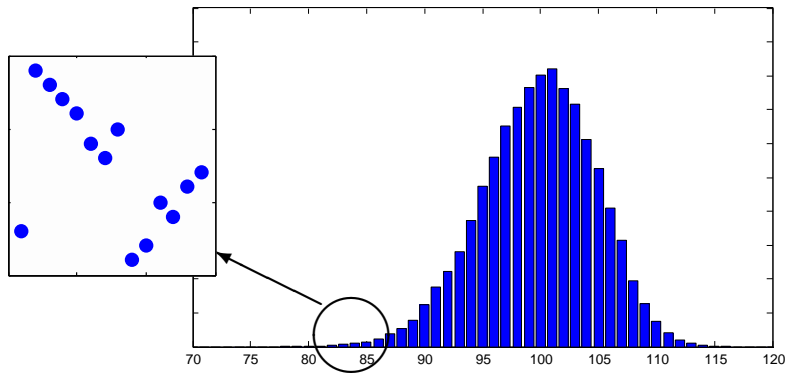
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



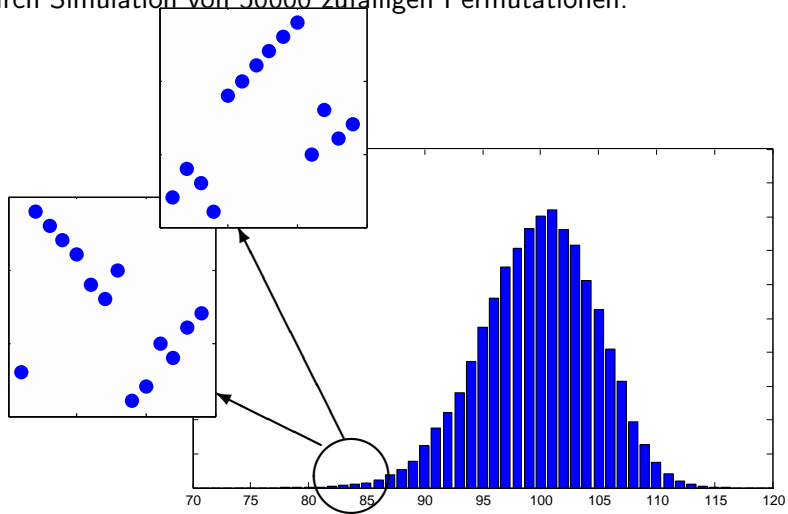
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



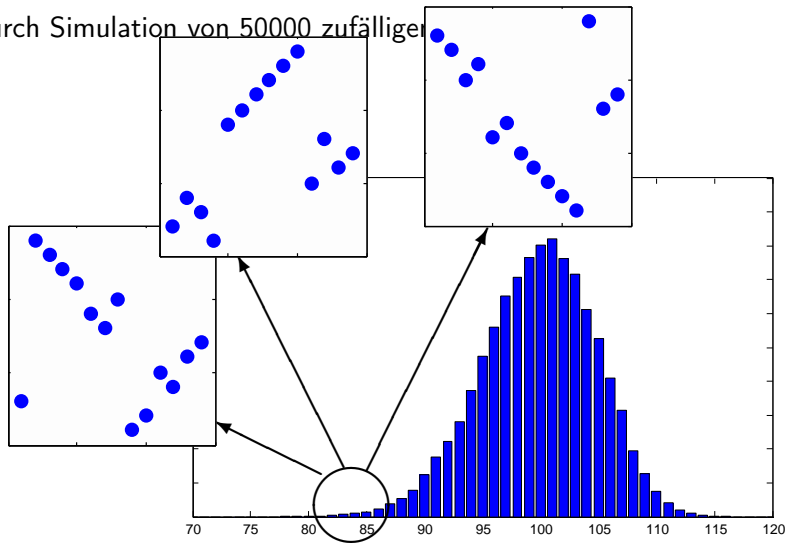
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



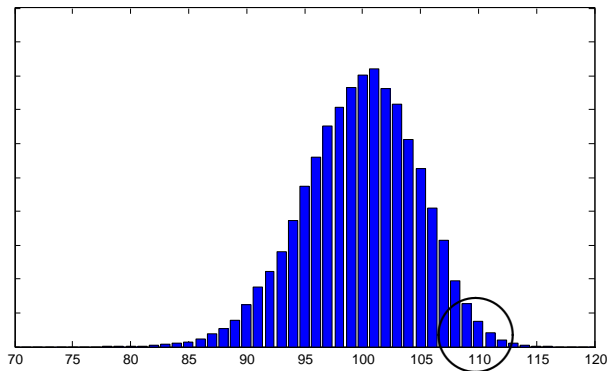
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälliger



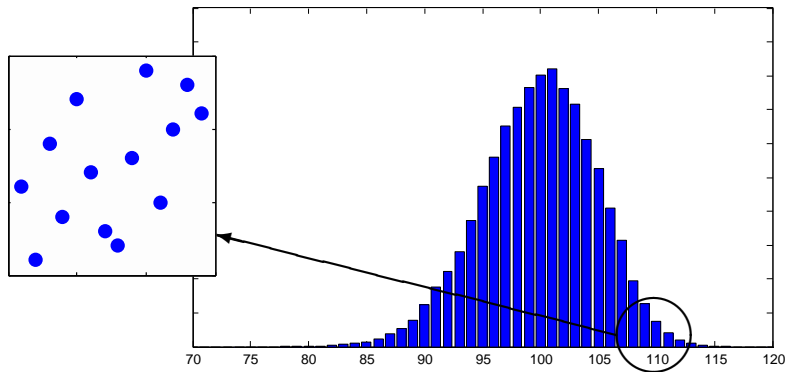
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



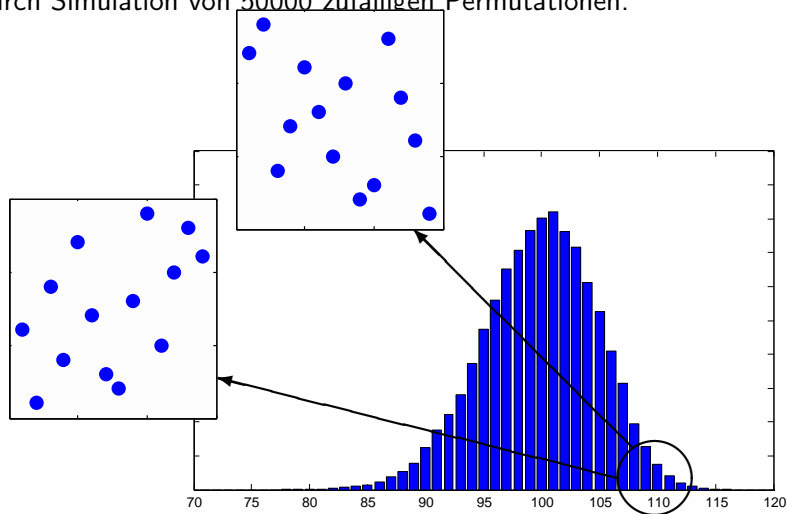
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



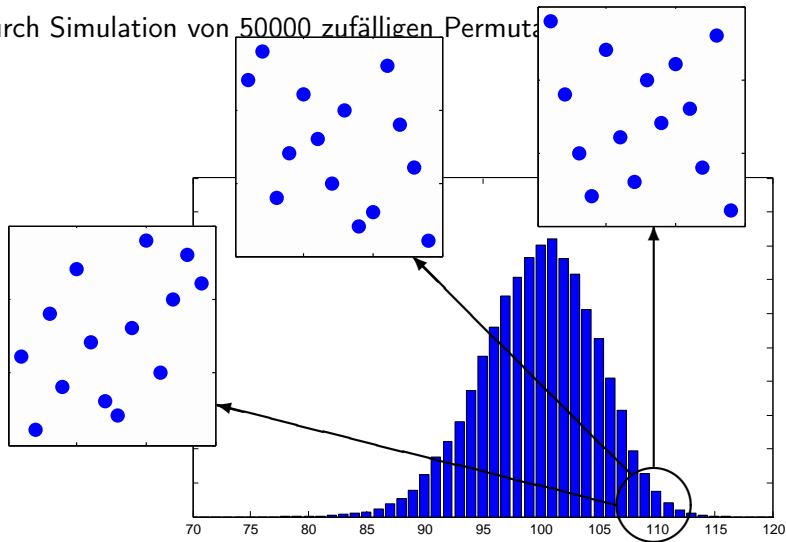
Verteilung von S_1 für $n = 14$

durch Simulation von 50000 zufälligen Permutationen.



Verteilung von S_1 für $n = 14$

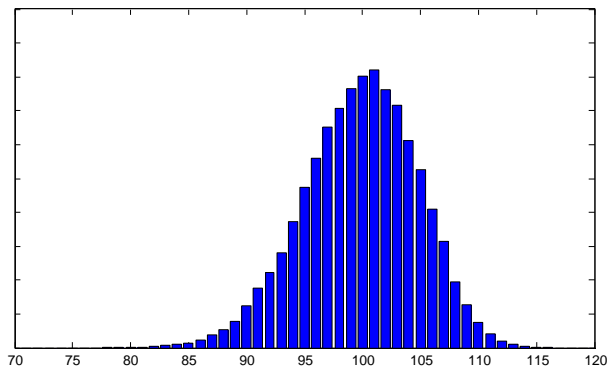
durch Simulation von 50000 zufälligen Permuta



Verteilung von S_1

H_0 : X und Y sind unabhängig

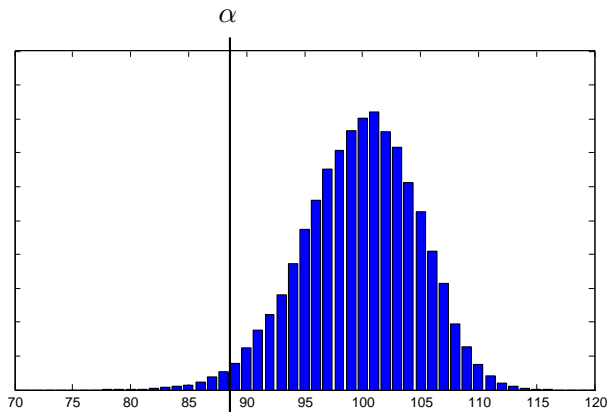
H_1 : X und Y sind abhängig



Verteilung von S_1

H_0 : X und Y sind unabhängig

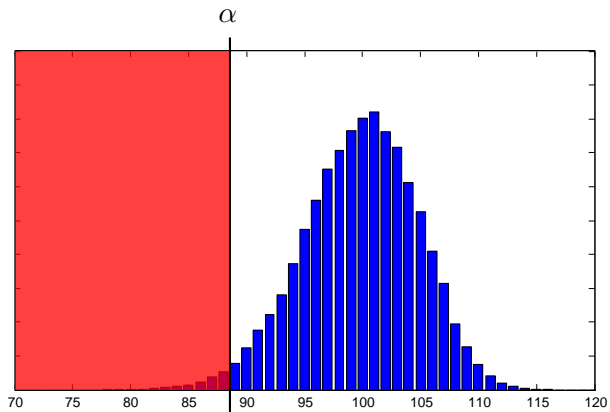
H_1 : X und Y sind abhängig



Verteilung von S_1

H_0 : X und Y sind unabhängig

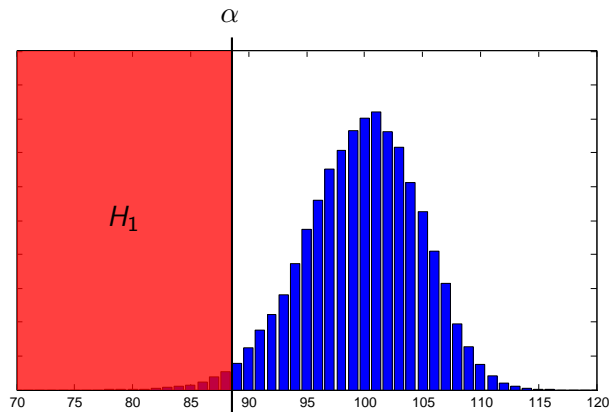
H_1 : X und Y sind abhängig



Verteilung von S_1

H_0 : X und Y sind unabhängig

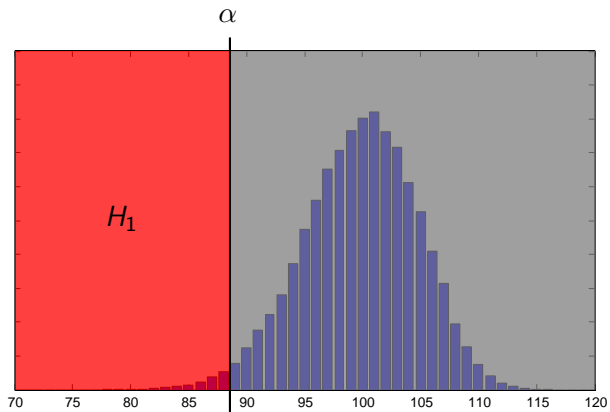
H_1 : X und Y sind abhängig



Verteilung von S_1

H_0 : X und Y sind unabhängig

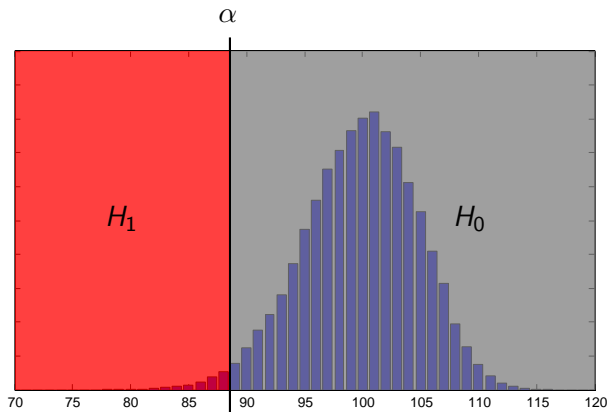
H_1 : X und Y sind abhängig

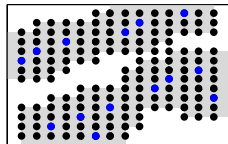


Verteilung von S_1

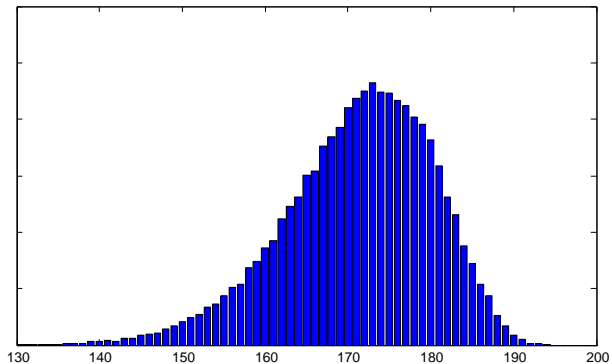
H_0 : X und Y sind unabhängig

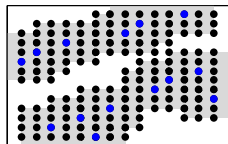
H_1 : X und Y sind abhängig



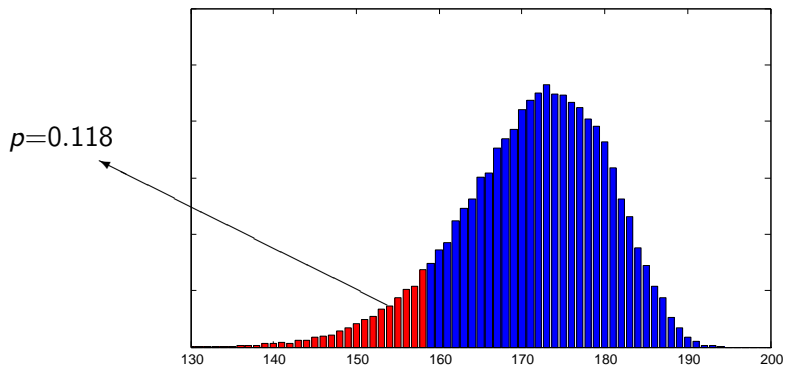
Verteilung von S_2 

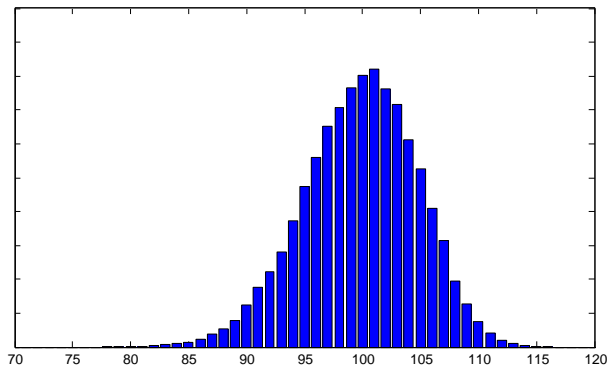
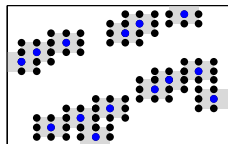
$$\rightarrow s_2 = 159$$

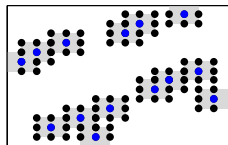


Verteilung von S_2 

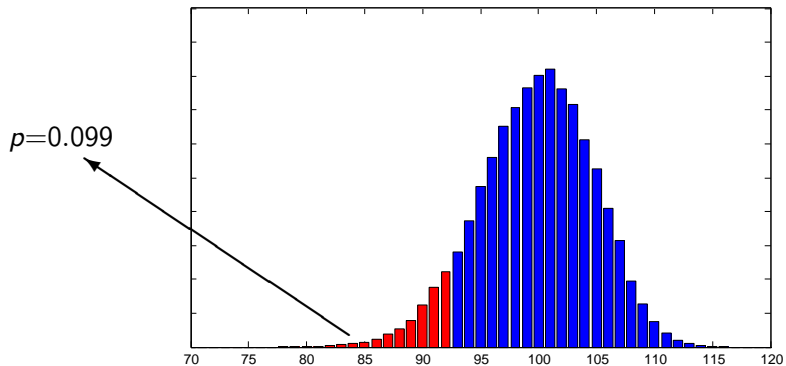
$$\rightarrow s_2 = 159$$



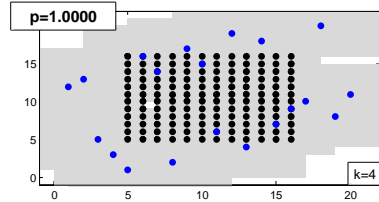
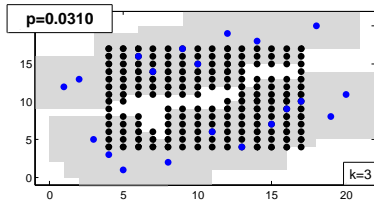
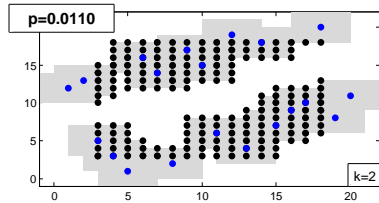
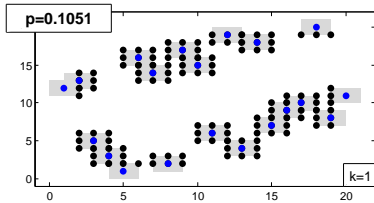
Verteilung von S_1 

Verteilung von S_1 

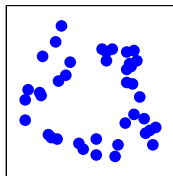
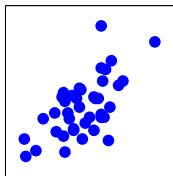
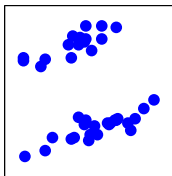
$$\rightarrow s_1 = 92$$



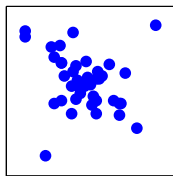
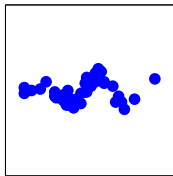
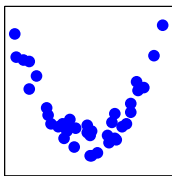
Die Qual der Wahl: Der Radius r

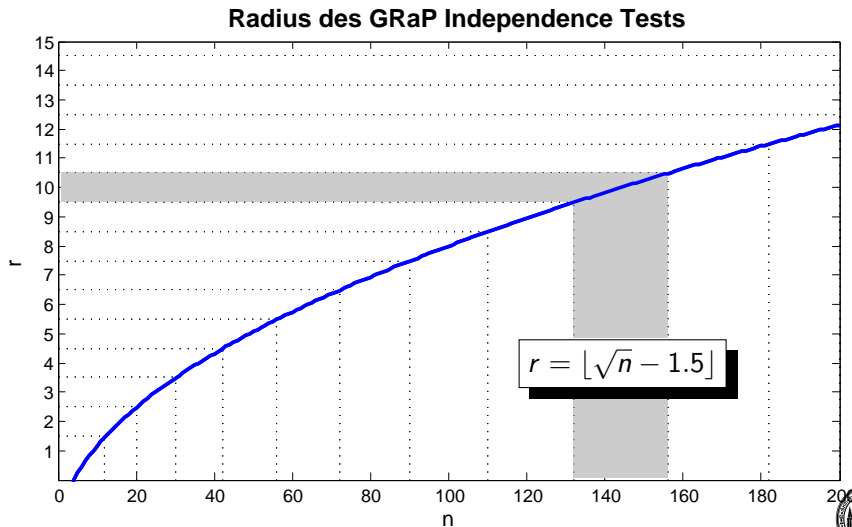


Die Qual der Wahl: Der Radius r



- Berechne die Teststärke von GRaP bei diesen Verteilungen für $n=2, \dots, 250$ und verschiedene $r=1, \dots, 20$
- Wähle den Radius der durchschnittlich die beste Teststärke liefert



Die Qual der Wahl: Der Radius r 

Berechnung des p-Wertes



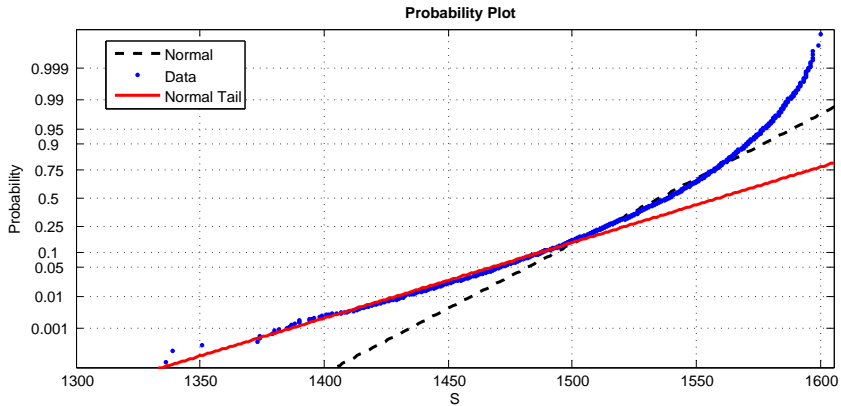
Quantilstabellen

- Berechnung von 1000er-Quantilen aus 10001 zufälligen Permutationen
- 50-fache Wiederholung
- Der Durchschnitt jedes 1000er-Quantils als Schätzung des wahren Quantils
- Fehlerabschätzung möglich

n	r	Quantile von S_r		
		0.05	0.025	0.01
10	1	57	55	54
15	2	172	167	162
20	2	271	265	258



Normalverteilungsapproximation



Power Analyse



Was ist ein guter Test?

- Test akzeptiert H_0 bei unabhängigen Merkmalen
- Test lehnt H_0 ab bei abhängigen Merkmalen



Was ist ein guter Test?

- Test akzeptiert H_0 bei unabhängigen Merkmalen
- Test lehnt H_0 ab bei abhängigen Merkmalen

- $P(H_1|H_0)$ Fehler 1.Art (α -Fehler, p -Wert)
- $P(H_0|H_1)$ Fehler 2.Art (β -Fehler)
- $P(H_0|H_0)$ Spezifität
- $P(H_1|H_1)$ Sensitivität



Untersuchte Verteilungsarten

- 1 Spezifität
- 2 Sensitivität



Untersuchte Verteilungsarten

- ① Spezifität
 - Gleichverteilung
 - Normalverteilung
- ② Sensitivität



Untersuchte Verteilungsarten

1 Spezifität

- Gleichverteilung
- Normalverteilung

2 Sensitivität

- Linear Distribution
- Linear Distribution Mix
- Normal Distribution Mix
- Circular Distribution



Untersuchte Verteilungsarten

1 Spezifität

- Gleichverteilung
- Normalverteilung

2 Sensitivität

- Linear Distribution
- Linear Distribution Mix
- Normal Distribution Mix
- Circular Distribution

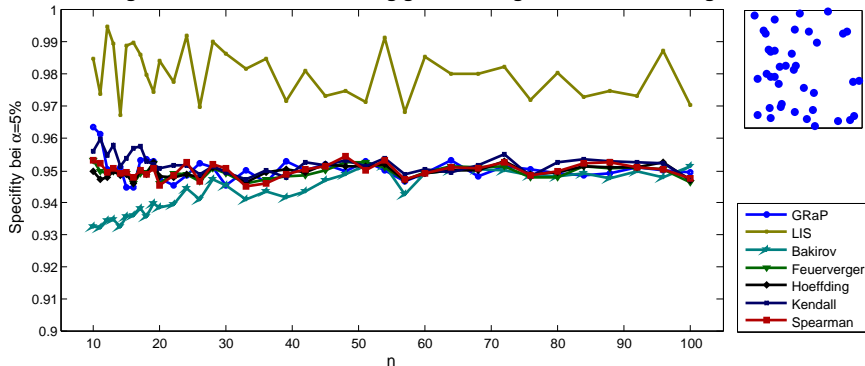
Schätzung der Teststärke

Berechnung der p-Werte je Verteilungsart und je Stichprobenumfang aus 10000 Stichproben.



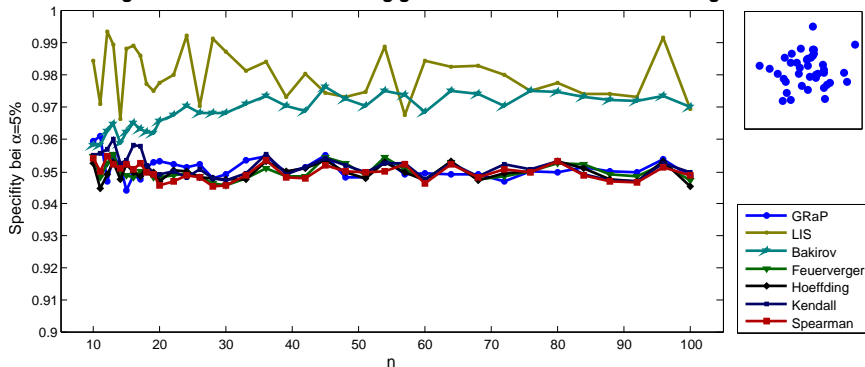
Spezifität

Einhaltung des α -Niveaus für unabhängige bivariate gleichverteilte Zufallsgrößen

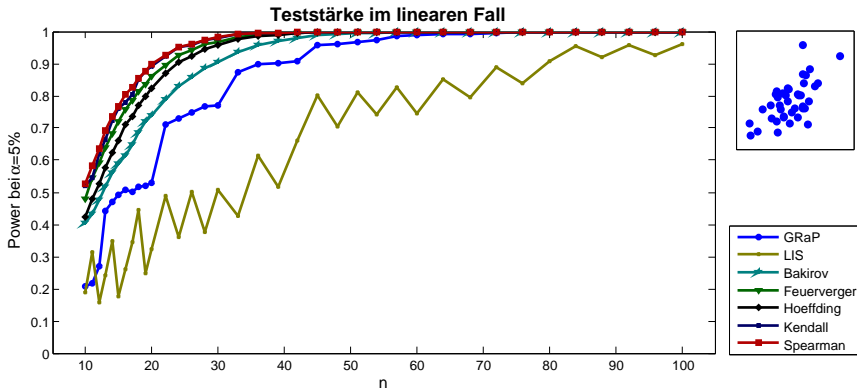


Spezifität

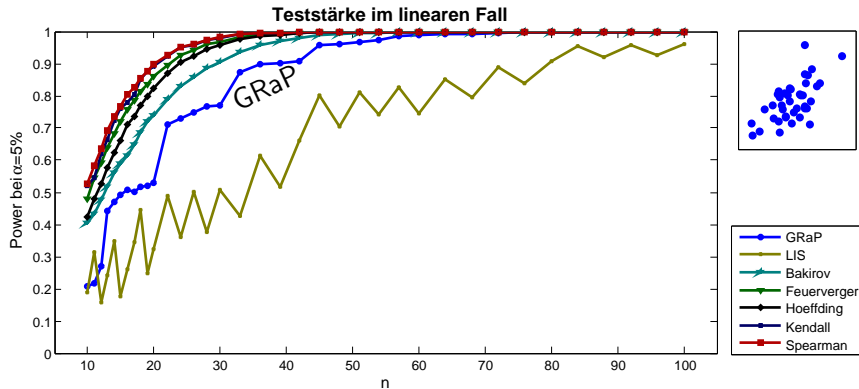
Einhaltung des α -Niveaus für unabhängige bivariate normalverteilte Zufallsgrößen



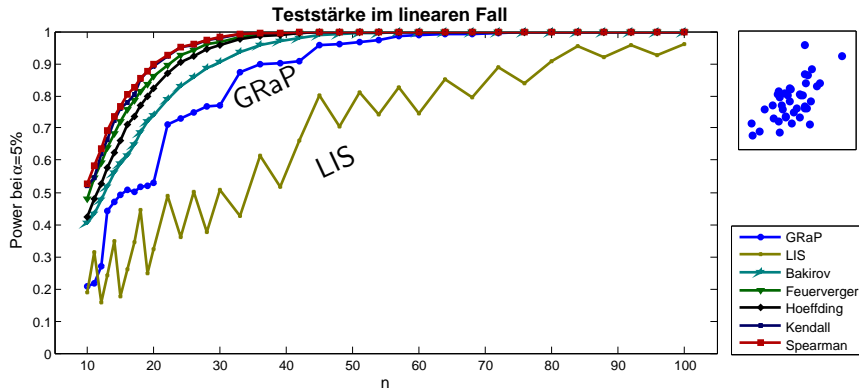
Sensitivität



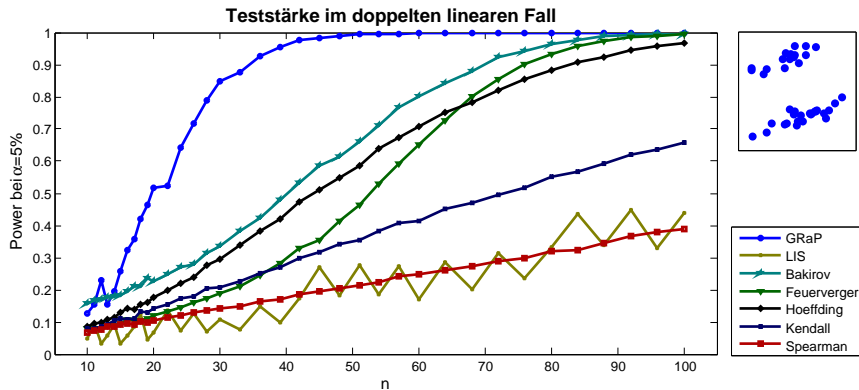
Sensitivität



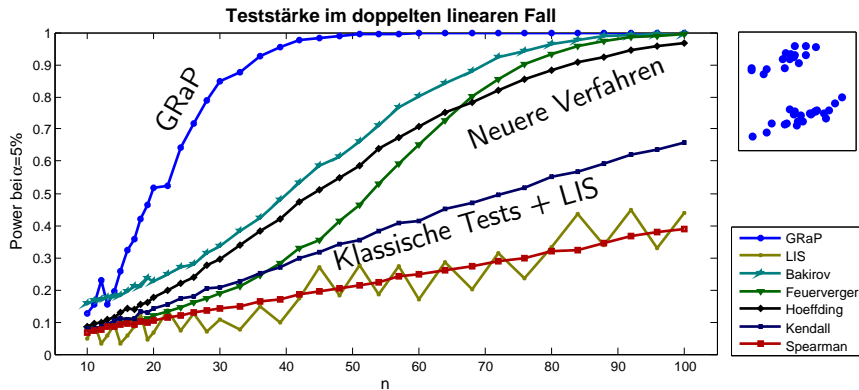
Sensitivität



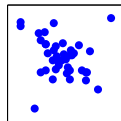
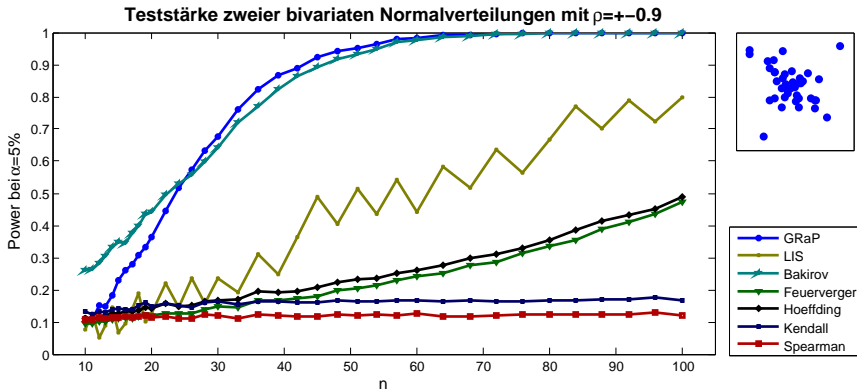
Power



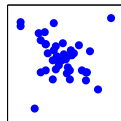
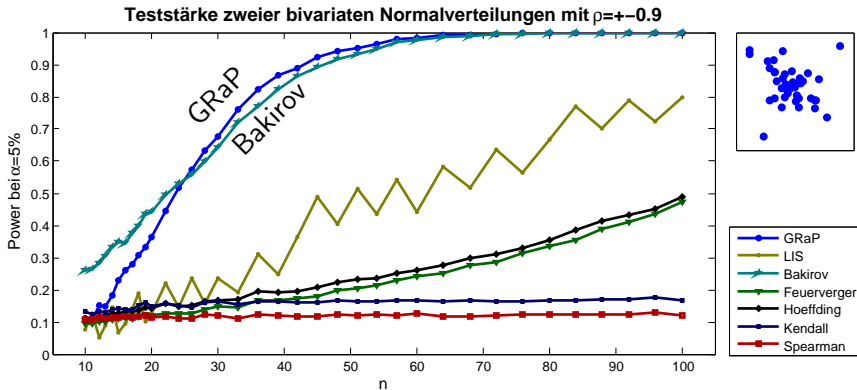
Power



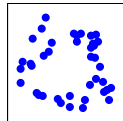
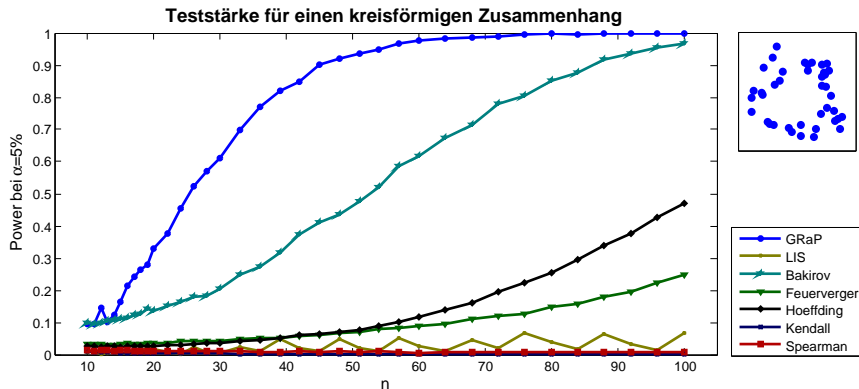
Power



Power



Power

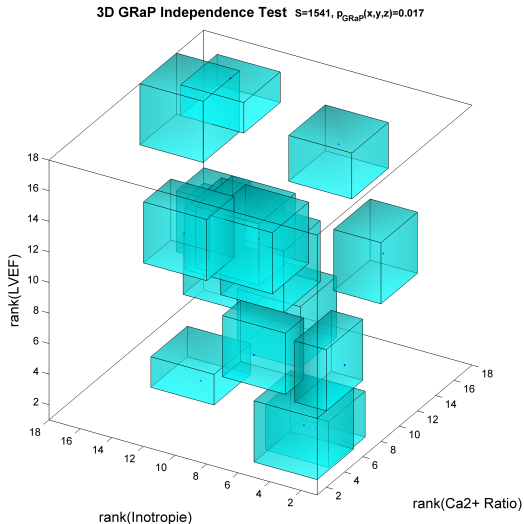
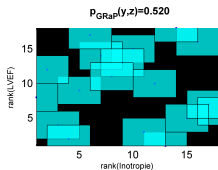
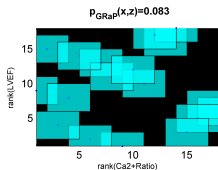
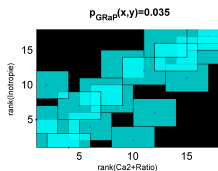


Beispiel



Immunadsorptionsverfahren

Universitätsmedizin Greifswald, vor IA



Zusammenfassung



Vor- und Nachteile

Vorteile

Nachteile



Vor- und Nachteile

Vorteile

- Hohe Teststärke bei nichtlinearen Zusammenhängen

Nachteile



Vor- und Nachteile

Vorteile

- Hohe Teststärke bei nichtlinearen Zusammenhängen
- Verständliche und leicht zu berechnende Teststatistik

Nachteile



Vor- und Nachteile

Vorteile

- Hohe Teststärke bei nichtlinearen Zusammenhängen
- Verständliche und leicht zu berechnende Teststatistik
- Bildung des p-Wertes durch Quantilstabellen oder Berechnung aus Normalverteilungen

Nachteile



Vor- und Nachteile

Vorteile

- Hohe Teststärke bei nichtlinearen Zusammenhängen
- Verständliche und leicht zu berechnende Teststatistik
- Bildung des p-Wertes durch Quantilstabellen oder Berechnung aus Normalverteilungen

Nachteile

- Testgröße nicht analytisch bestimmt



Diskussion





Vielen Dank für Ihre Aufmerksamkeit!



Referenzen







ausgewählte Literatur

- 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. 
- C. Spearman (1904), *The Proof and Measurement of Association between Two Things*, The American Journal of Psychology, Vol. 15, pp. 72-101. 
- 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. 
- M. G. Kendall (1938), *A New Measure Of Rank Correlation*, Biometrika, Vol. 30, pp. 81-93. 






ausgewählte Literatur

- G. A. Barnard (1947), *Significance Tests for 2×2 Tables*, Biometrika, Vol. 34, pp. 123-138. 
- W. Hoeffding (1948), *A Non-Parametric Test of Independence*, The Annals of Mathematical Statistics, Vol. 19, pp. 546-557. 
- RCE1962R. 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. 
- F1993A. Feuerverger (1993), *A Consistent Test for Bivariate Dependence*, International Statistical Review / Revue Internationale de Statistique, Vol. 61, pp. 419-433. 



ausgewählte Literatur

- 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. 
- M. Matilla-García, M. R. Marín (2008), *A non-parametric independence test using permutation entropy*, Journal of Econometrics, Vol. 144, pp. 139-155. 
- J. E. García, V. A. González-López (2009), *A Nonparametric Independence Test using Random Permutations*, Preprint, arXiv:0908.2794v2. 

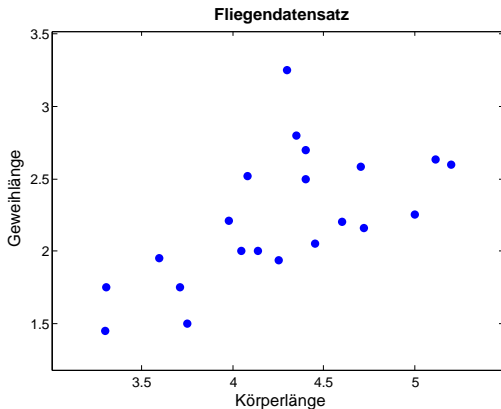


Appendix



Fliegendatensatz

www.stat.uni-muenchen.de/service/datenarchiv/fliegen/fliegen.html



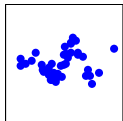
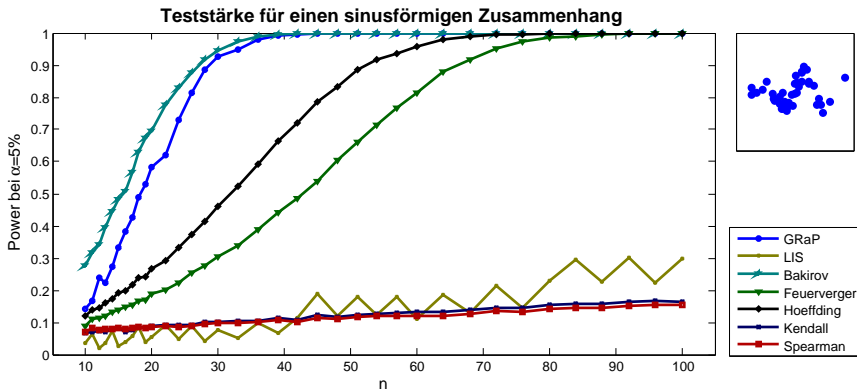
Pearsonkorrelation: $r = 0.63$
Spearmankorrelation: $r_s = 0.67$
Kendalls Tau: $\tau = 0.47$

p-Werte:

Pearson = 0.0024
Spearman = 0.0009
Chi2-Ind. = 0.0046
Fisher = 0.0073
Kendall = 0.0030
Quadranten = 0.1615
Barnard = 0.0037
Hoeffding = 0.0036
Feuerverger = 0.0010
Bakirov = 0.0070
GarciaMarin = 0.8100
LIS = 0.4961
GRaP = 0.0020



Power



Power

