A new Independence Test for continuous variables

Marcus Vollmer

Department of Mathematics and Computer Science
University of Greifswald

COMPUTING & STATISTICS (ERCIM'11)

12-19-2011

1. Problem statement
2. GRaP Independence Test
3. Calculation of the p-value
4. Power Analysis
5. Example

Problem statement

- Continuous variables X, Y

e.g. conifer: X - tree height
Y - length of treetop

- Empirical values $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

Problem statement

- Continuous variables X, Y

e.g. conifer: X - tree height
Y - length of treetop

- Empirical values $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$

- Testing hypothesis:

$$\begin{aligned} \mathrm{H}_0 &: \quad X \text{ and } Y \text{ are independent} \\ \mathrm{H}_1 &: \quad X \text{ and } Y \text{ are not independent} \end{aligned}$$

A look on Independence Tests

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

A look on Independence Tests

1895	Pearsons r	Linear or monotone Dependencies
1904	Spearmans ρ	
1938	Kendalls τ	
1922	χ^2-Test	
1925	Fisher-Yates-Test	
1945	Barnards CSM-Test	
1948	Hoeffdings D-Test	
1993	Feuerverger-Test	
2004	Bakirovs I_n	
2009	LIS-Test	
2010	GRaP-Test	

A look on Independence Tests

1895	Pearsons r	Linear or monotone Dependencies
1904	Spearmans ρ	
1938	Kendalls τ	
1922	χ^2-Test	Categorical / Frequency-based
1925	Fisher-Yates-Test	
1945	Barnards CSM-Test	
1948	Hoeffdings D-Test	
1993	Feuerverger-Test	
2004	Bakirovs I_n	
2009	LIS-Test	
2010	GRaP-Test	

A look on Independence Tests

1895	Pearsons r	Linear or monotone Dependencies
1904	Spearmans ρ	
1938	Kendalls τ	
1922	χ^2-Test	Categorical / Frequency-based
1925	Fisher-Yates-Test	
1945	Barnards CSM-Test	
1948	Hoeffdings D-Test	Ordinal / Various Dependencies
1993	Feuerverger-Test	
2004	Bakirovs I_n	
2009	LIS-Test	
2010	GRaP-Test	

A look on Independence Tests

1895	Pearsons r	Linear or monotone Dependencies
1904	Spearmans ρ	
1938	Kendalls τ	
1922	χ^2-Test	Categorical / Frequency-based
1925	Fisher-Yates-Test	
1945	Barnards CSM-Test	
1948	Hoeffdings D-Test	Independence of continuous variables as ordinal problem, because for monotone $\phi : \mathbb{R} \to \mathbb{R}$, we have: X, Y indep. $\Leftrightarrow \phi(X), Y$ indep.
1993	Feuerverger-Test	
2004	Bakirovs I_n	
2009	LIS-Test	
2010	GRaP-Test	

A look on Independence Tests

Hoeffding, Feuerverger, Bakirov

- using difference between the joint distribution function $F(x, y)$ and marginal distributions $F(x), F(y)$
- Test statistic: $||F(x, y) - F(x)F(y)||$

A look on Independence Tests

Hoeffding, Feuerverger, Bakirov

- using difference between the joint distribution function $F(x, y)$ and marginal distributions $F(x), F(y)$
- Test statistic: $||F(x, y) - F(x)F(y)||$

LIS, GRaP

- using permutations

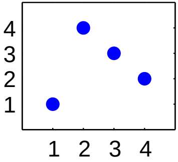
$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ y_1 & y_2 & \dots & y_n \end{pmatrix} \xrightarrow[\text{sorted by x}]{\text{ranks of } x,y} \begin{pmatrix} 1 & 2 & \dots & n \\ \sigma(1) & \sigma(2) & \dots & \sigma(n) \end{pmatrix} = \sigma \in S_n$$

$\sigma : (1, 2, \dots, n) \to \{1, 2, \dots, n\}$

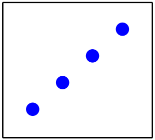
- Independence ⇒ Every Permutation has the same prob. to appear

Test statistic for permutations?

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



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

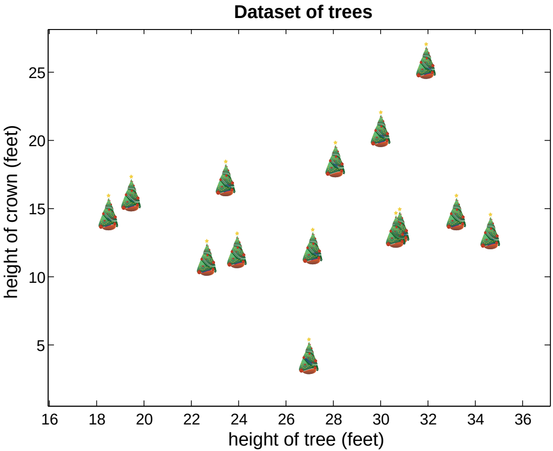


should be in the rejection region

- **LIS-Test**: maximal length of increasing subsequences
- **GRaP**: "Geometry of Random Permutations"
  - by Christoph Bandt and Marcus Vollmer
  - uses the geometric arrangement of the corresponding permutation of the sample
  - for small sample sizes

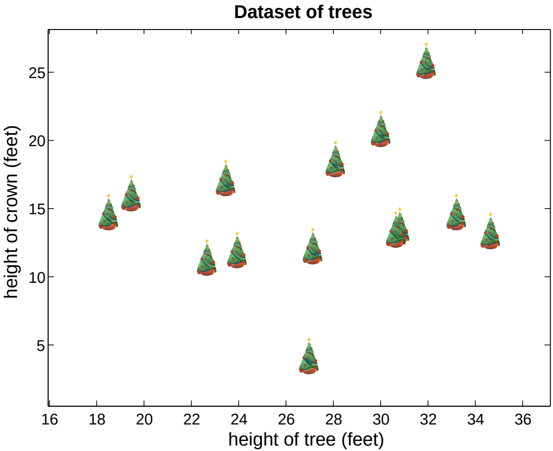
Conifers in Southwest Oregon

$n = 14$



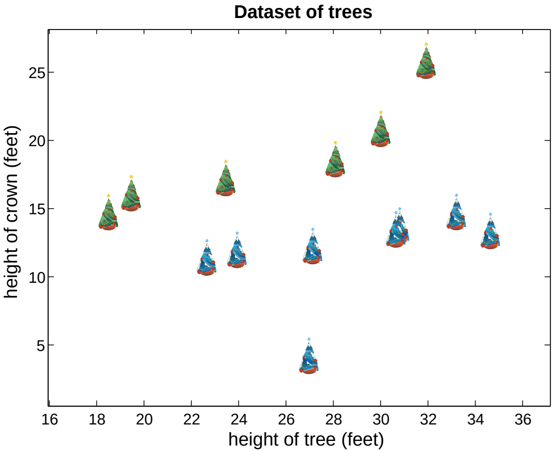
Conifers in Southwest Oregon

$n = 14$



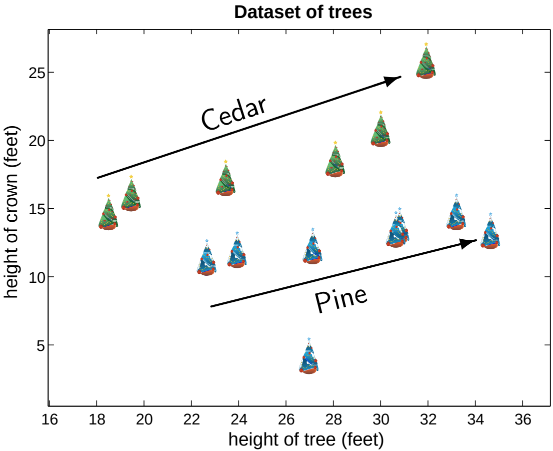
Conifers in Southwest Oregon

n=14



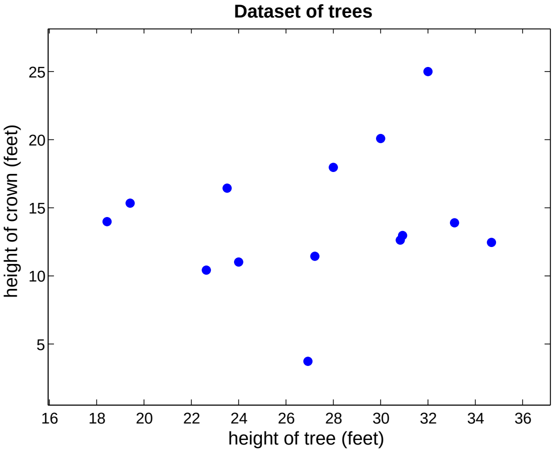
Conifers in Southwest Oregon

$n{=}14$



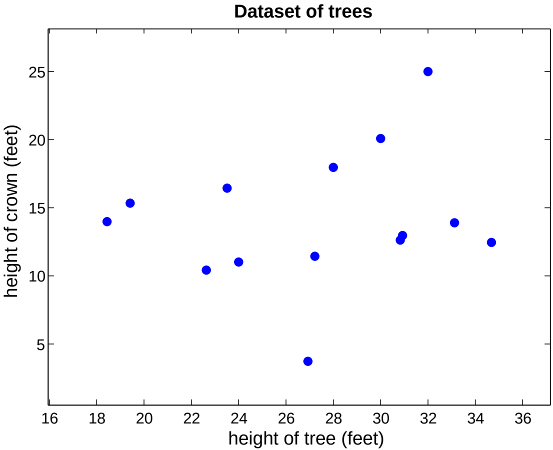
GRaP Independence Test

n=14



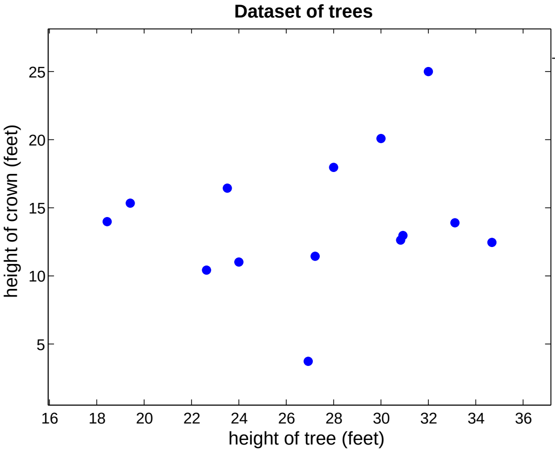
GRaP Independence Test

n=14



GRaP Independence Test

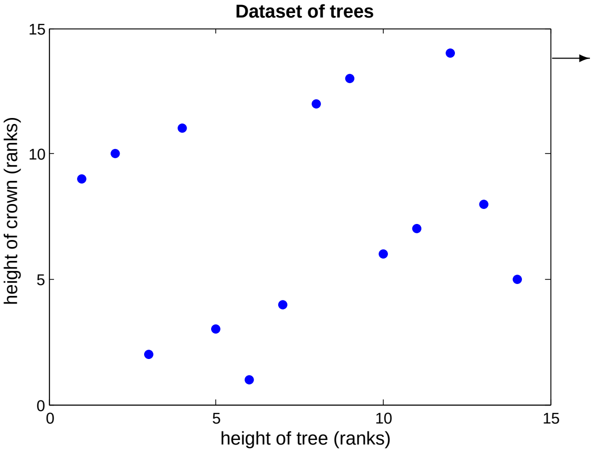
$n=14$



→ Rank transform

GRaP Independence Test

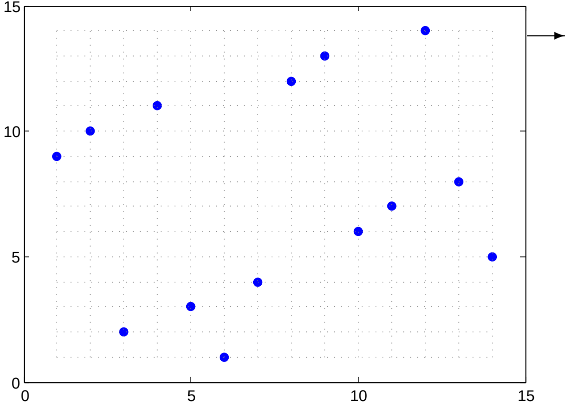
n=14



→ Rank transform

GRaP Independence Test

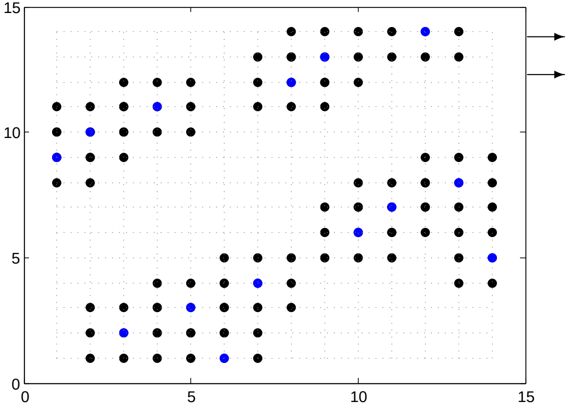
n=14



→ Rank transform

GRaP Independence Test

n=14 , r=1



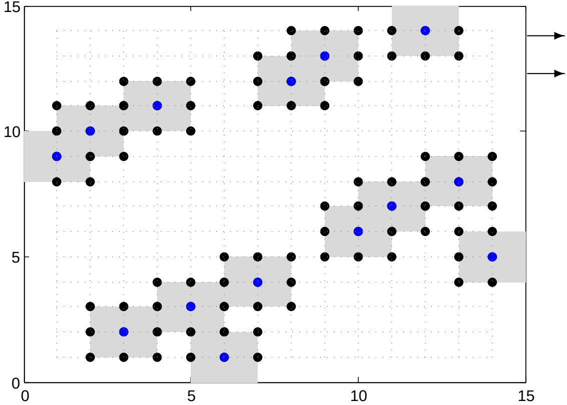
Rank transform

Counting

neighbor coordinates

GRaP Independence Test

$n=14$, $r=1$



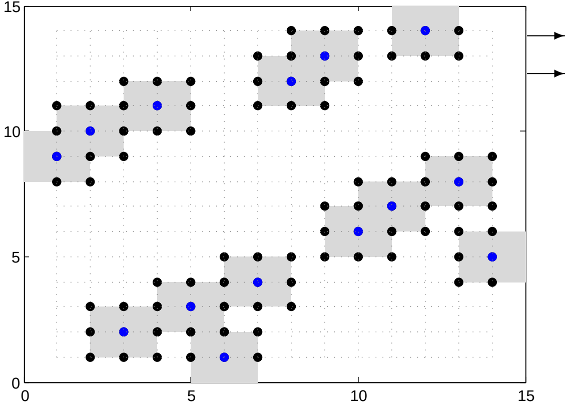
Rank transform

Counting

neighbor coordinates

GRaP Independence Test

$n=14$, $r=1$



Rank transform

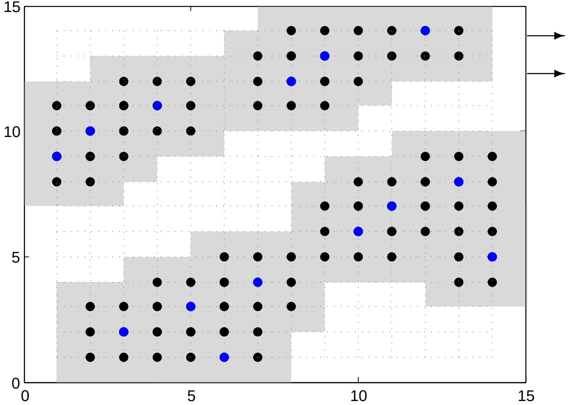
Counting

neighbor coordinates

Number: $s_1=92$

GRaP Independence Test

$n=14$, $r=2$



Rank transform

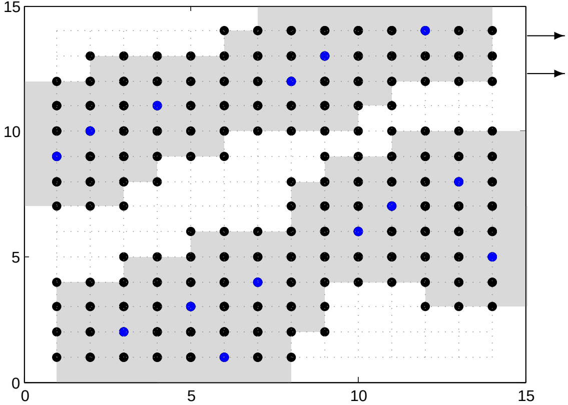
Counting

neighbor coordinates

Number: $s_1=92$

GRaP Independence Test

$n = 14$, $r = 2$



Rank transform

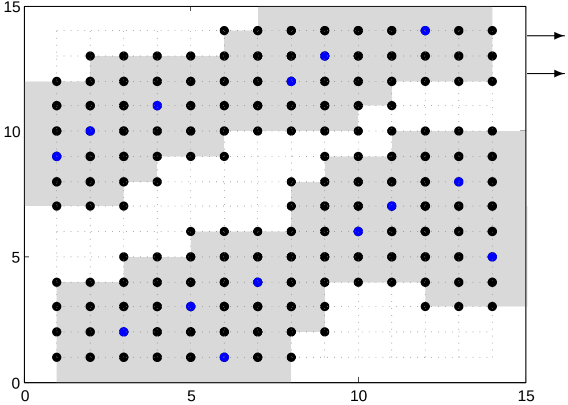
Counting

neighbor coordinates

Number: $s_1 = 92$

GRaP Independence Test

$n=14$, $r=2$



Rank transform

Counting

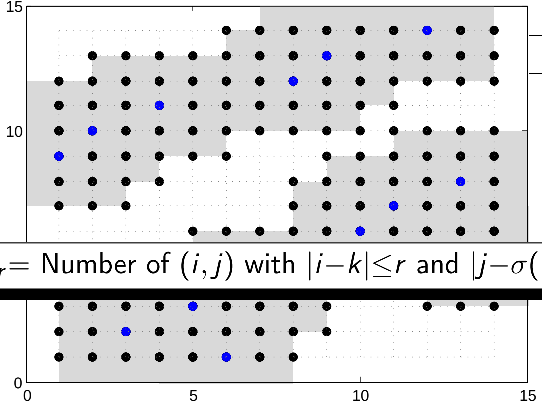
neighbor coordinates

Number: $s_1=92$

$s_2=159$

GRaP Independence Test

$n = 14$, $r = 2$



→ Rank transform

→ Counting neighbor coordinates

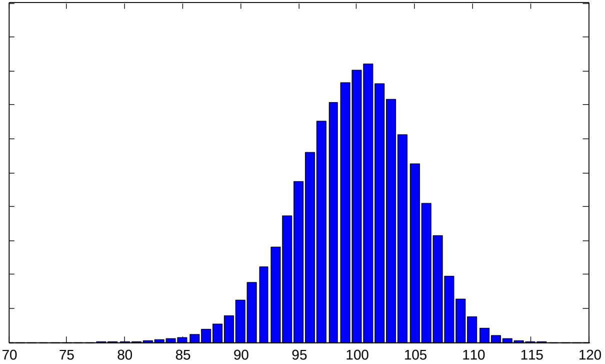
Number: $s_1 = 92$

$s_2 = 159$

$S_r =$ Number of (i,j) with $|i - k| \leq r$ and $|j - \sigma(k)| \leq r$ for some k

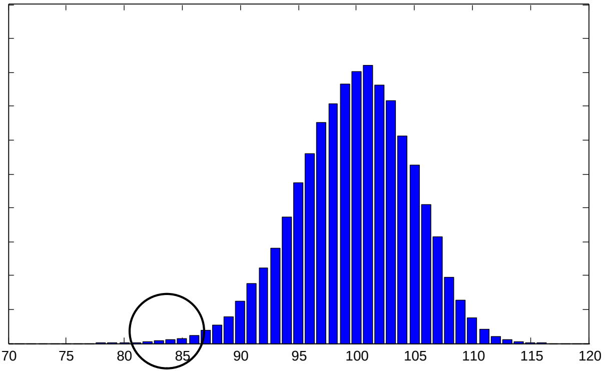
Distribution of S_1 for $n = 14$

using 50000 random permutations



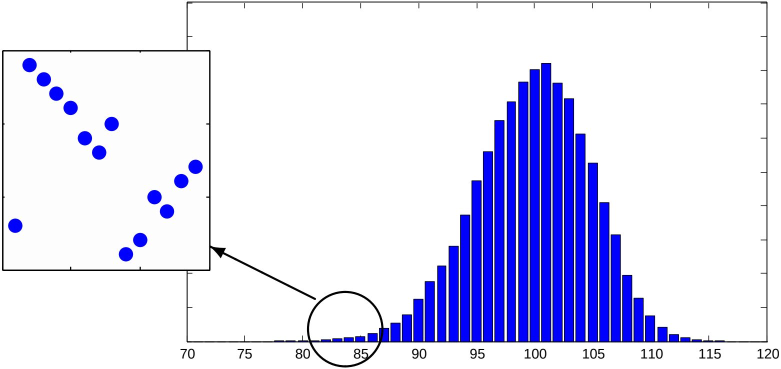
Distribution of S_1 for $n = 14$

using 50000 random permutations



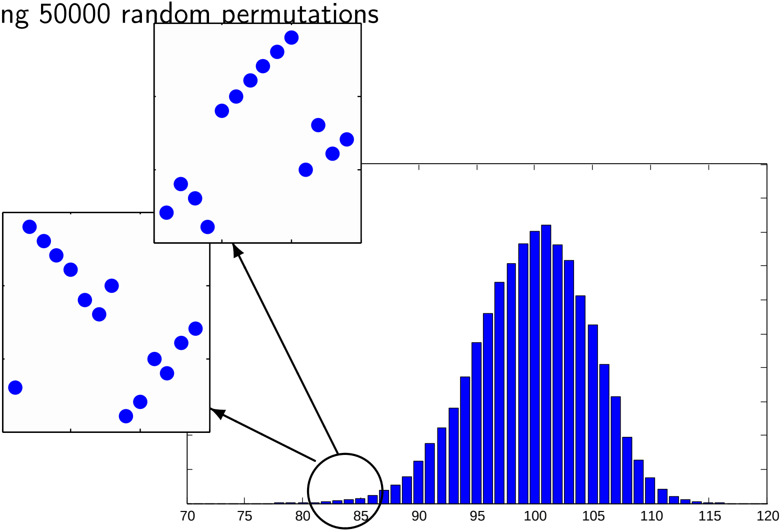
Distribution of S_1 for $n = 14$

using 50000 random permutations



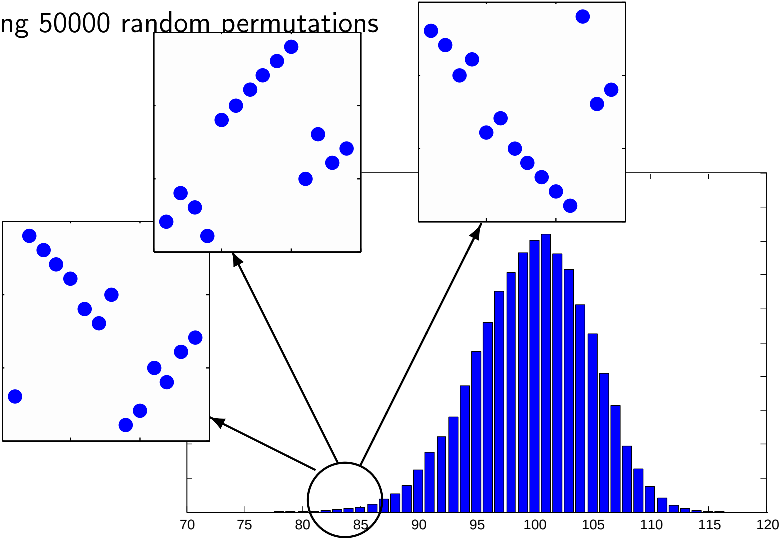
Distribution of S_1 for $n = 14$

using 50000 random permutations



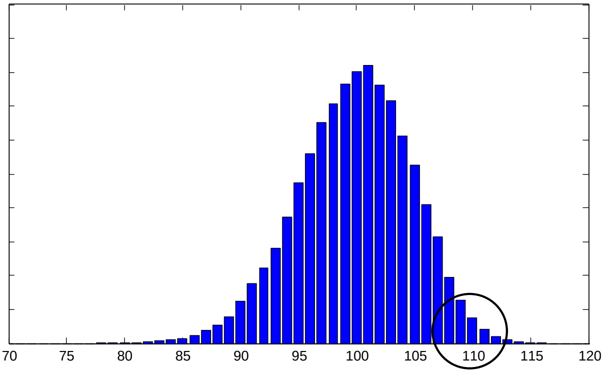
Distribution of S_1 for $n = 14$

using 50000 random permutations



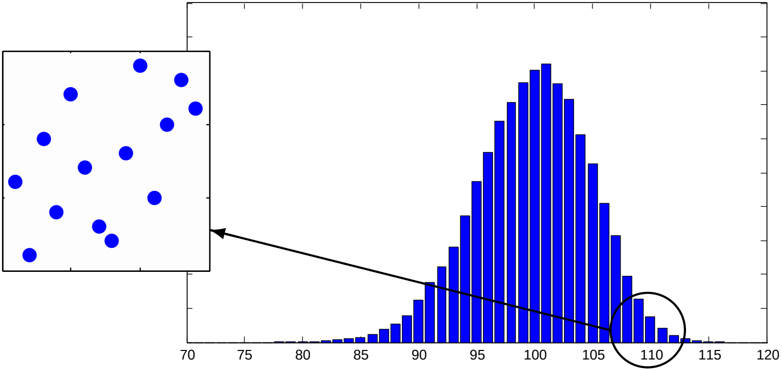
Distribution of S_1 for $n = 14$

using 50000 random permutations



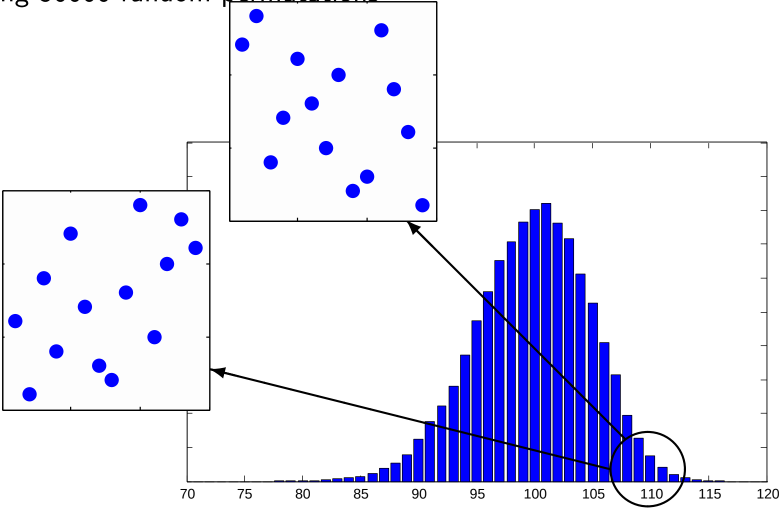
Distribution of S_1 for $n = 14$

using 50000 random permutations



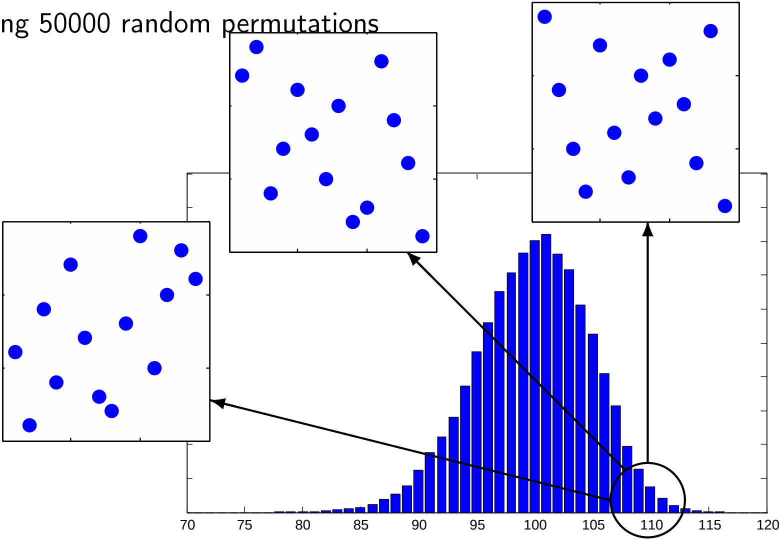
Distribution of S_1 for $n = 14$

using 50000 random permutations



Distribution of S_1 for $n = 14$

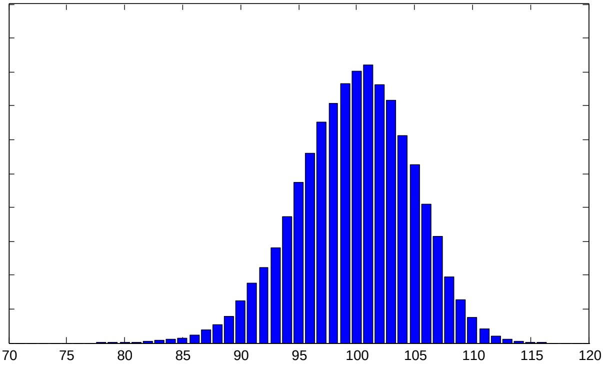
using 50000 random permutations



Distribution of S_1

H_0 : X and Y are independent

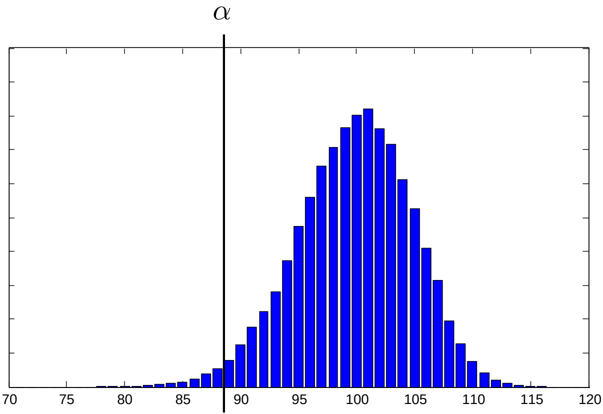
H_1 : X and Y are dependent



Distribution of S_1

H_0 : X and Y are independent

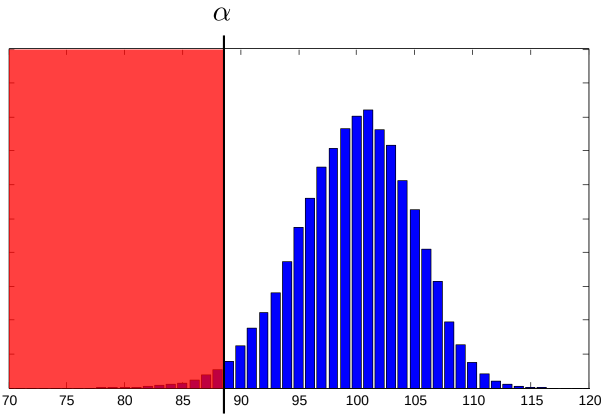
H_1 : X and Y are dependent



Distribution of S_1

H_0 : X and Y are independent

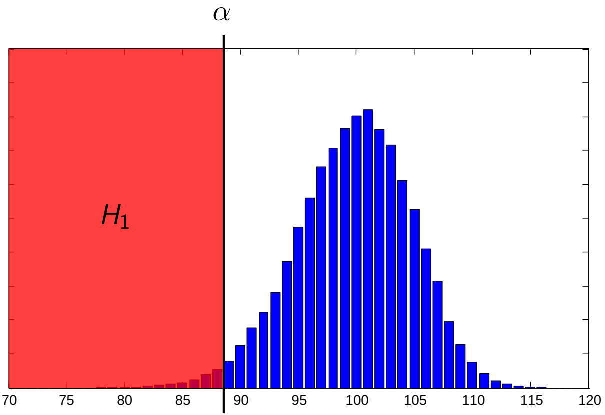
H_1 : X and Y are dependent



Distribution of S_1

H_0 : X and Y are independent

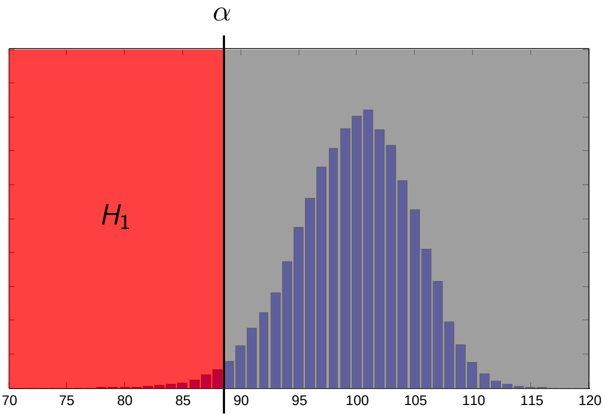
H_1 : X and Y are dependent



Distribution of S_1

H_0 : X and Y are independent

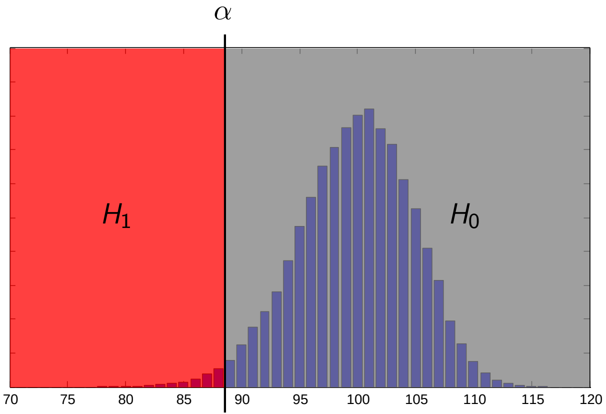
H_1 : X and Y are dependent



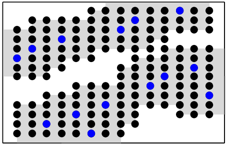
Distribution of S_1

H_0 : X and Y are independent

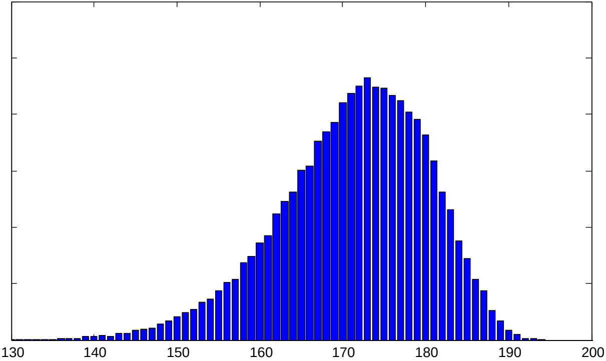
H_1 : X and Y are dependent



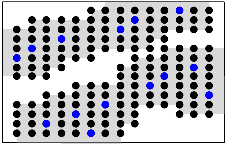
Distribution of S_2



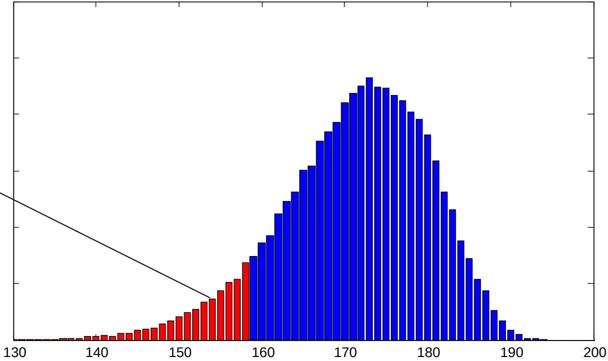
$\longrightarrow \quad s_2 = 159$



Distribution of S_2

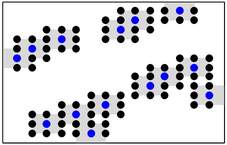


$\longrightarrow$ $s_2 = 159$

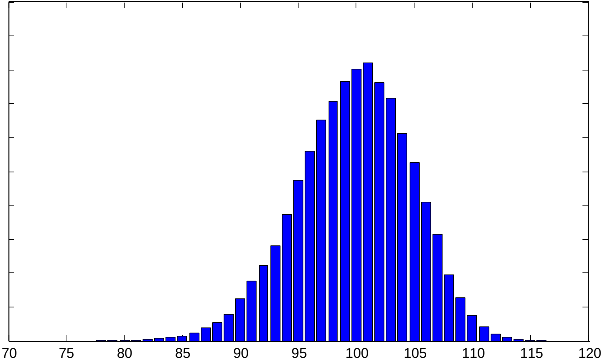


$p = 0.118$

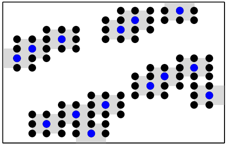
Distribution of S_1



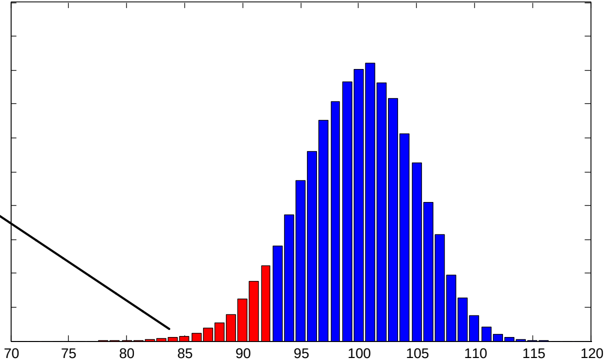
$\longrightarrow$ $s_1 = 92$



Distribution of S_1

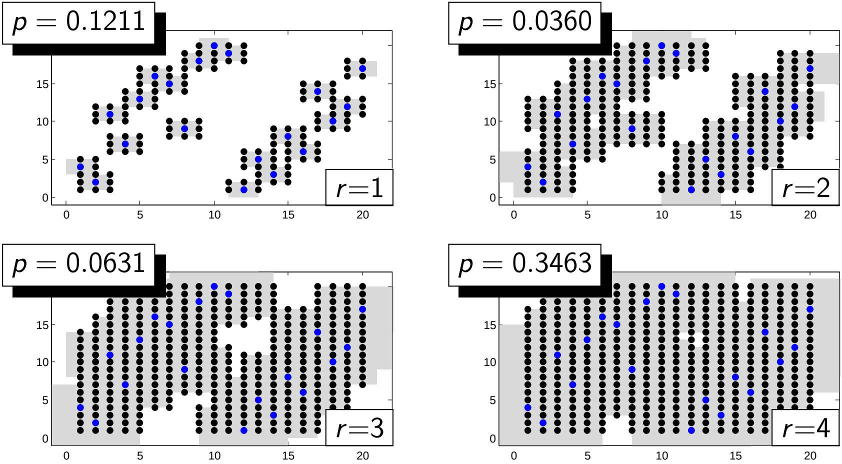


→ $s_1=92$

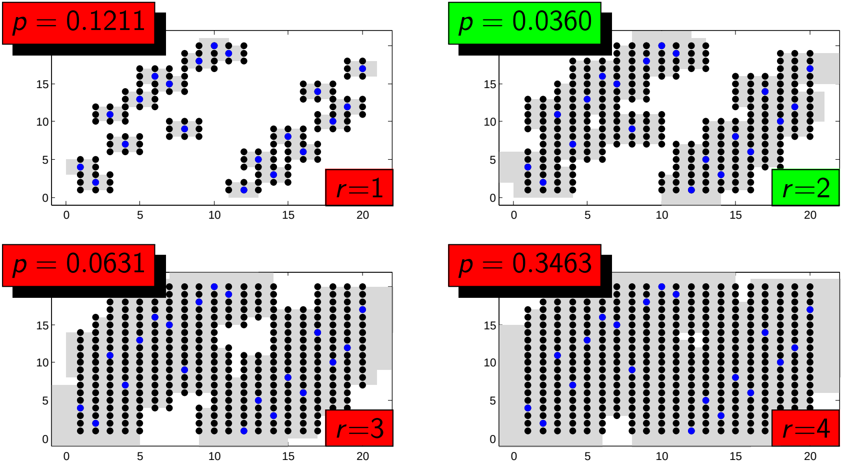


$p=0.099$

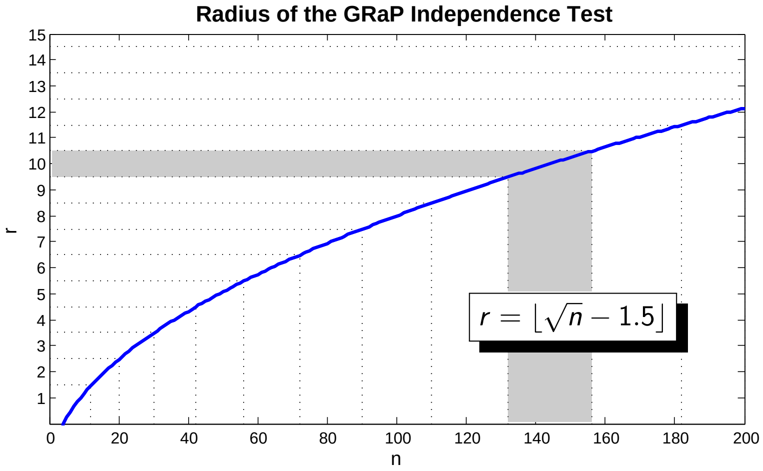
Spoilt for choice: The radius r



Spoilt for choice: The radius *r*



Spoilt for choice: The radius r



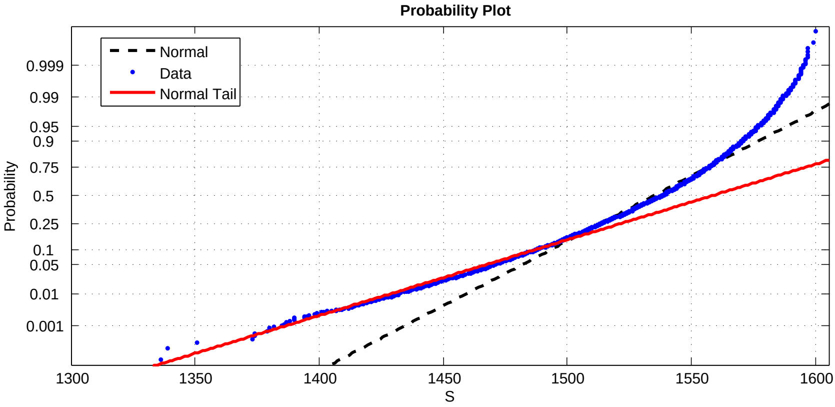
$$r = \lfloor \sqrt{n} - 1.5 \rfloor$$

Table of quantiles

- Calculation of permilles out of 10001 random permutations
- The mean of 50 repetitions as estimation of the distribution
- Error estimation possible

		p-quantile of S_r		
n	**r**	**0.05**	**0.025**	**0.01**
10	**1**	57	55	54
15	**2**	172	167	162
20	**2**	271	265	258

Approximation by normal distributions



Analysed distribution types

1. Specificity $P(H_0|H_0)$
   - Uniform
   - Normal
2. Sensitivity $P(H_1|H_1)$
   - Linear
   - Linear distribution mix
   - Normal distribution mix
   - Circular

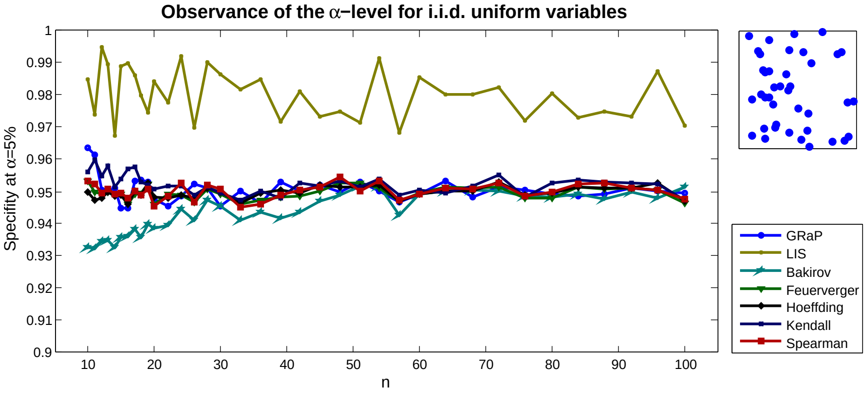
Analysed distribution types

1. Specificity $P(H_0|H_0)$
   - Uniform
   - Normal
2. Sensitivity $P(H_1|H_1)$
   - Linear
   - Linear distribution mix
   - Normal distribution mix
   - Circular

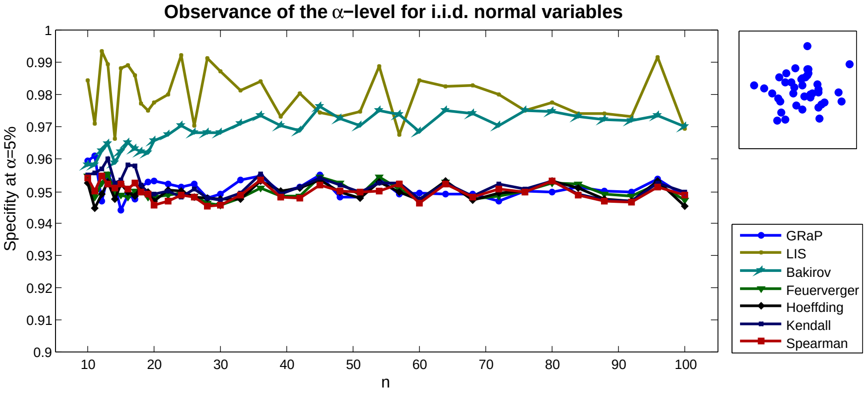
Estimation of the power efficiency

Evaluation of 10000 p-values for every kind of distr. and every sample size.

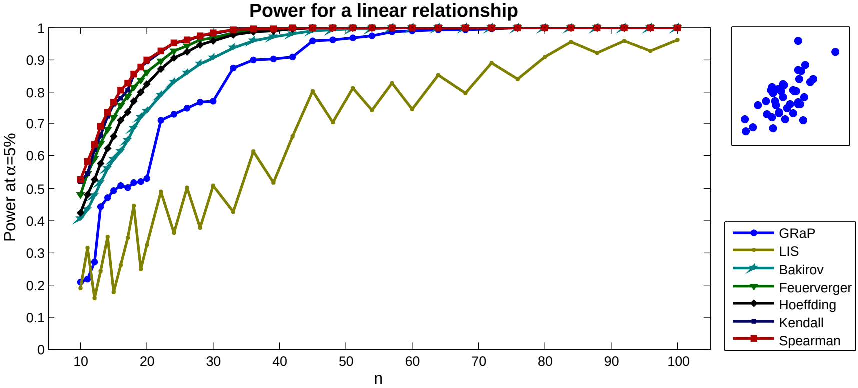
Specificity



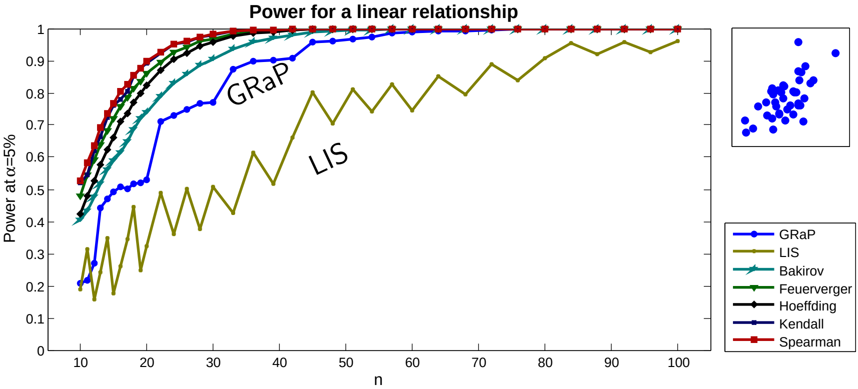
Specificity



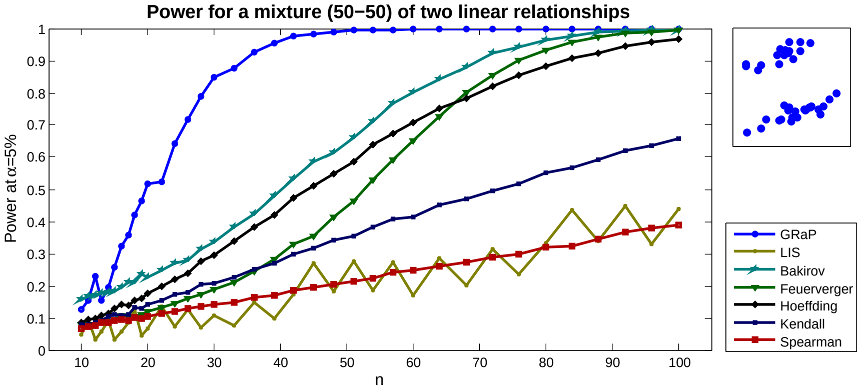
Sensitivity



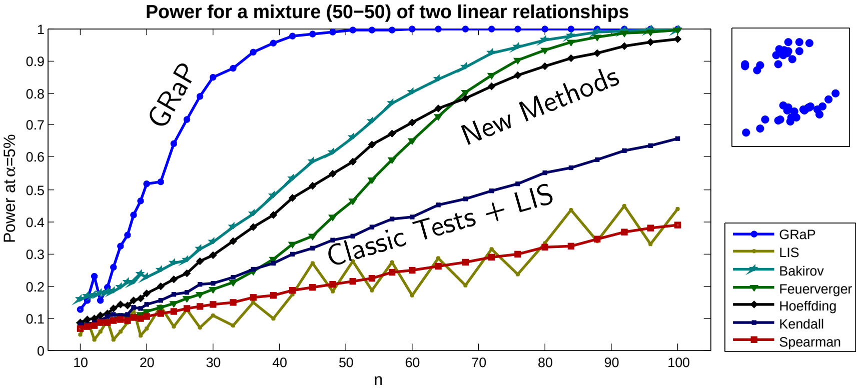
Sensitivity



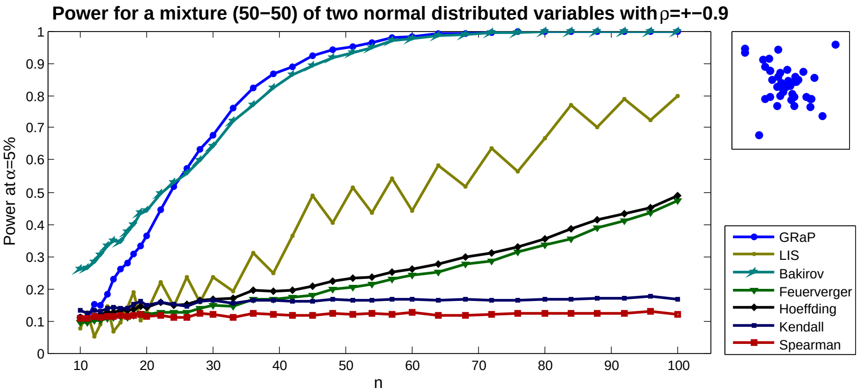
Power


Power for a mixture (50−50) of two linear relationships
Power at α=5%
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
70
80
90
100
n
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman

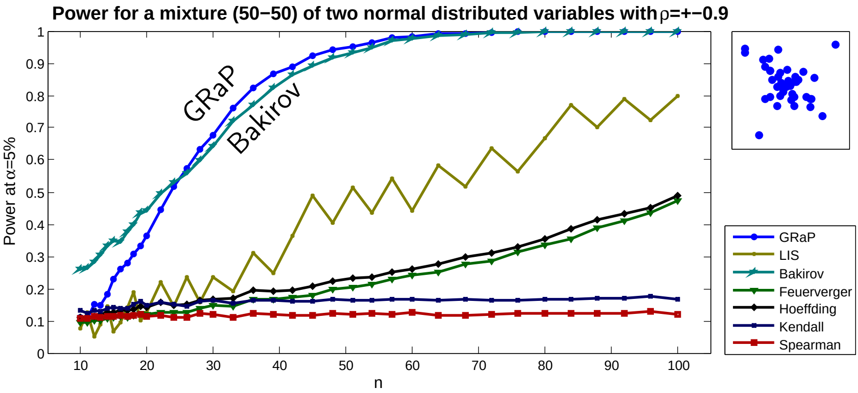
Power


Power for a mixture (50−50) of two linear relationships
Power at α=5%
n
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
70
80
90
100
GRaP
New Methods
Classic Tests + LIS
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman

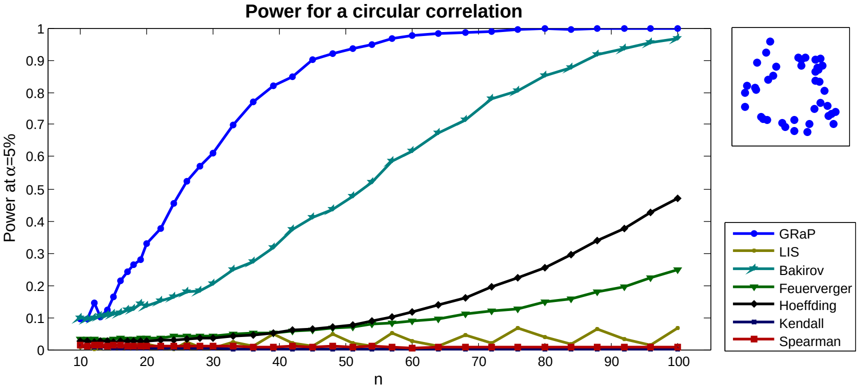
Power


Power for a mixture (50−50) of two normal distributed variables with ρ=+−0.9
Power at α=5%
n
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
70
80
90
100
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman

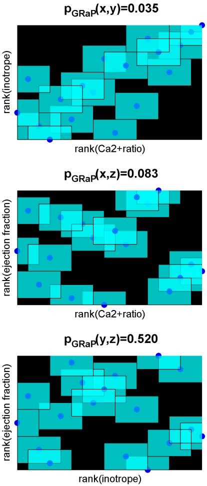
Power

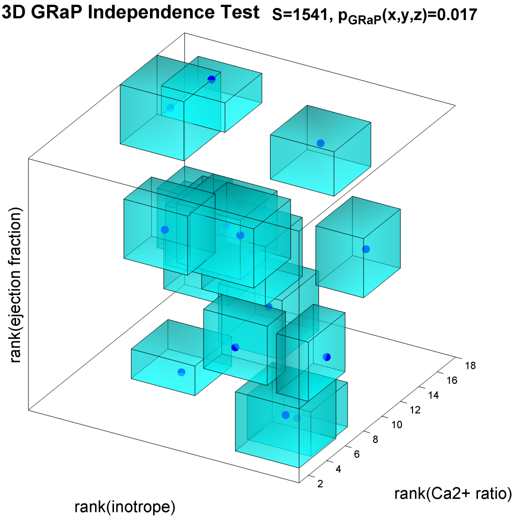

Power for a mixture (50−50) of two normal distributed variables with ρ=+−0.9
Power at α=5%
n
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
70
80
90
100
GRaP
Bakirov
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman

Power

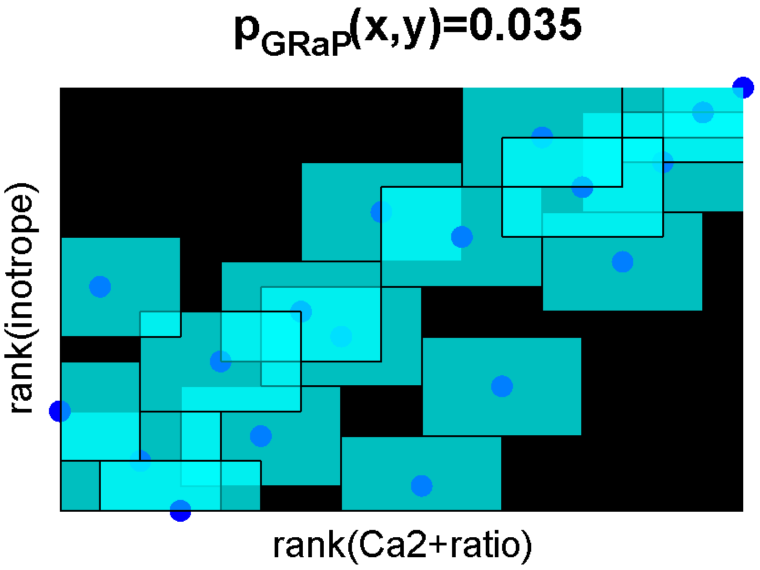

Power for a circular correlation
Power at α=5%
n
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman

Patient with autoimmune diseases





Patient with autoimmune diseases



Pros and cons

Advantages

Disadvantage

Pros and cons

Advantages

- Powerful in nonlinear relations
- Comprehensible test statistic which is easy to compute
- P-values estimated by Table of Quantiles or through normal distributions

Disadvantage

Pros and cons

Advantages

- Powerful in nonlinear relations
- Comprehensible test statistic which is easy to compute
- P-values estimated by Table of Quantiles or through normal distributions

Disadvantage

- Test statistic not analytically defined

Discussion

Thank you for your kind attention!

For further reading

- 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.

For further reading

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.

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.

A. Feuerverger (1993), *A Consistent Test for Bivariate Dependence*,
International Statistical Review / Revue Internationale de Statistique,
Vol. 61, pp. 419-433.

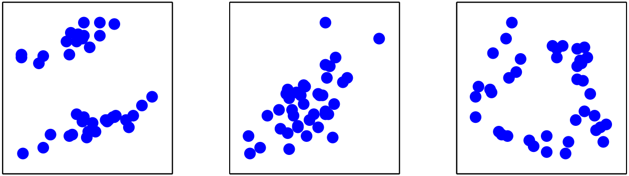
For further reading

- 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.

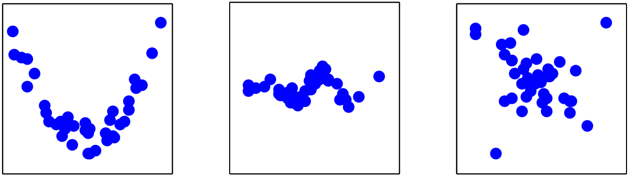
- M. Matilla-García, M. R. Marín (2008), *A non-parametric independence test using permutation entropy*,
  Journal of Econometrics, Vol. 144, pp. 139-155.

- J. E. García, V. A. González-López (2009), *A Nonparametric Independence Test using Random Permutations*,
  Preprint, arXiv:0908.2794v2.

Spoilt for choice: The radius r

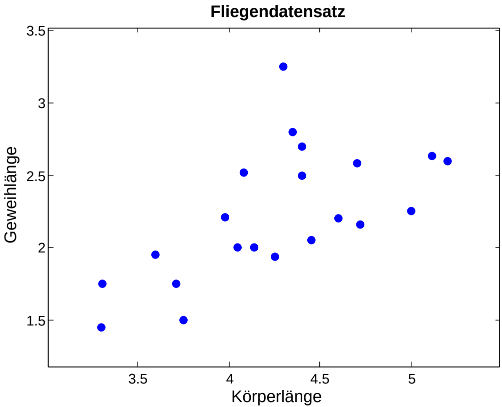


- For $n \in [2, 250]$ and $r \in [1, 20]$ we compute the power of GRaP for these types of distributions
- We choose the radius $r(n)$ which provides the best power (in average of dist.)



Dataset of flies

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



Pearsonkorrelation: r = 0.63

Spearmankorrelation: rs = 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

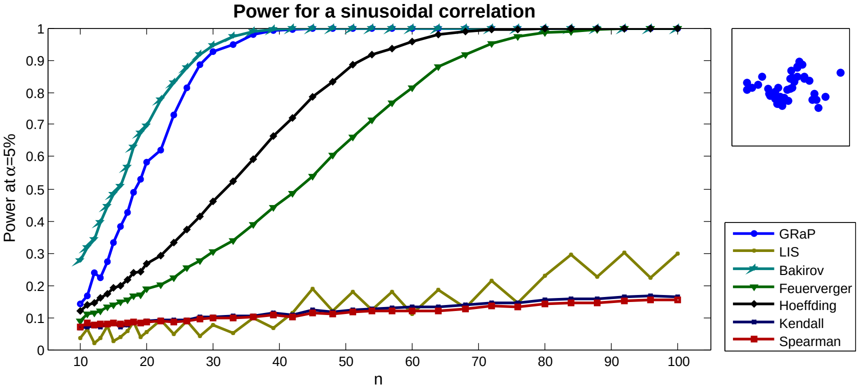
Bakirov = 0.0070

GarciaMarin = 0.8100

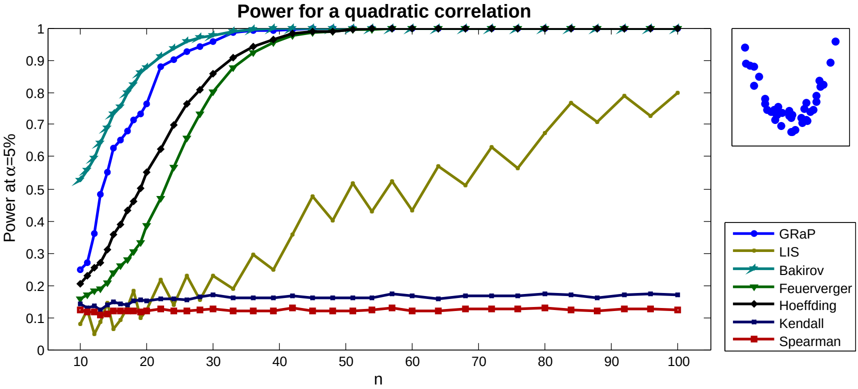
LIS = 0.4961

GRaP = 0.0020

Power


Power for a sinusoidal correlation
Power at α=5%
n
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
70
80
90
100
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman

Power


Power for a quadratic correlation
Power at α=5%
n
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
10
20
30
40
50
60
70
80
90
100
GRaP
LIS
Bakirov
Feuerverger
Hoeffding
Kendall
Spearman