非線形な相関関係の評価指標および推定手法

A Measure of Nonlinear Correlation with an Estimation Algorithm

鈴木 了太
suzuki@ef-prime.com
株式会社 ef-prime

2025年9月10日
2025年度 統計関連学会連合大会

背景

- 多変量のデータ行列が分析対象
- 強い関係を持つ変数の組を探索したい
  - 予測：目的変数Yと関係の強い説明変数X_i
  - 知識発見：強い関係をもつ変数の組(X_i, X_j)

既存手法

- 相関係数
  - 量的データ間の（線形な）相関関係
  - 例：積率相関係数、順位相関係数など
  - 非線形な関係を検出可能な手法もある
    - 距離相関（distance correlation; Székely et al, 2007）
    - MIC（maximal information coefficient; Reshef et al., 2011）

- 連関係数
  - 質的データ間の独立性
  - 例：クラメールのV、ピアソンのϕ係数など

要求

- 線形のみならず非線形な関係も評価
- データ型や分布を問わず、統一的に適用可能
  - 連続、離散、あるいはその混合
  - 外れ値や欠損を含む
  - 分布が非対称、裾が重い、多峰的

要求（できれば）

- 測定単位を問わない（変換に対して不変）
- データや分析手法の前提知識が不要
- 計算コストが低く、大規模データにも適用可能
- 統計的推論が可能（仮説検定、信頼区間の構成）

方針：従属性の評価

完全従属

- YがXに**完全従属**する（complete dependence; Lancaster, 1963）
  - XによるYの条件付き分布が一点分布になること
  - $P(Y = g(X)) = 1$となる可測関数gが存在

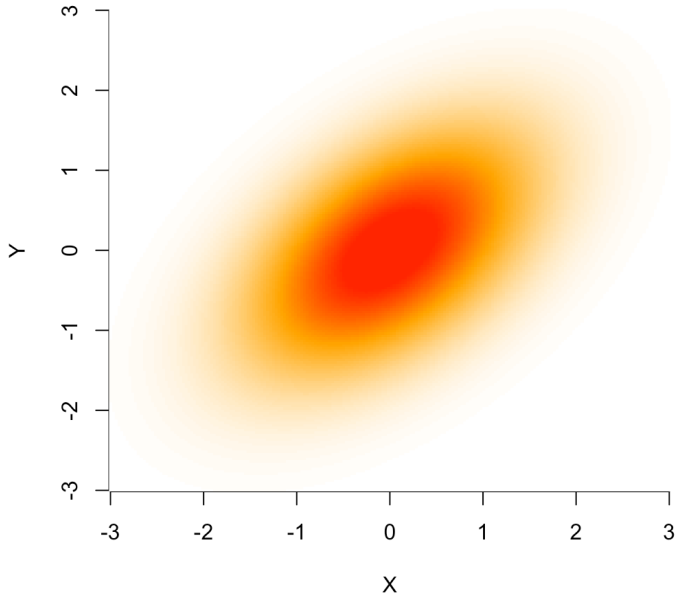
指標の設計方針

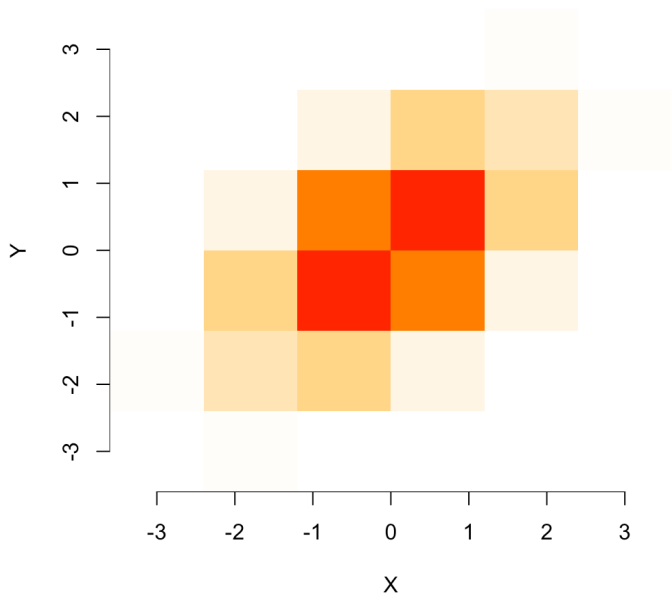
- XとYが独立のとき0
- YがXに（あるいは相互に）完全従属するとき1

> 従属性の指標が満たすべき性質はTasena & Dhompongsa (2016)などで議論されているが、当面は素朴な定義を用いる

着想：連続変数の離散化

連続変数の値域を分割し、離散変数として評価する





$$f(x, y)$$

$$p(i, j) = P(X \in \mathcal{X}_i, Y \in \mathcal{Y}_j)$$

$$X \in \mathcal{X},\ Y \in \mathcal{Y}$$

$$\mathcal{X} = \sqcup_{i=1}^{k} \mathcal{X}_i,\ \mathcal{Y} = \sqcup_{j=1}^{k} \mathcal{Y}_j$$

相互情報量（離散の場合）

- XとYが独立のとき0
- YがXに完全従属するときYの情報量$H(Y)$

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}\left[\log \frac{p(x,y)}{p_X(x)p_Y(y)}\right] \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)}
\end{aligned}
$$

$$
0 \leq I(X;Y) \leq H(Y)
$$

$$
H(Y) = -\sum_y p_Y(y) \log p_Y(y)
$$

不確実性係数（Uncertainty Coefficient）

相互情報量を目的変数の情報量で基準化

$$U(Y|X) = \frac{I(X;Y)}{H(Y)}$$

$$0 \leq U(Y|X) \leq 1$$

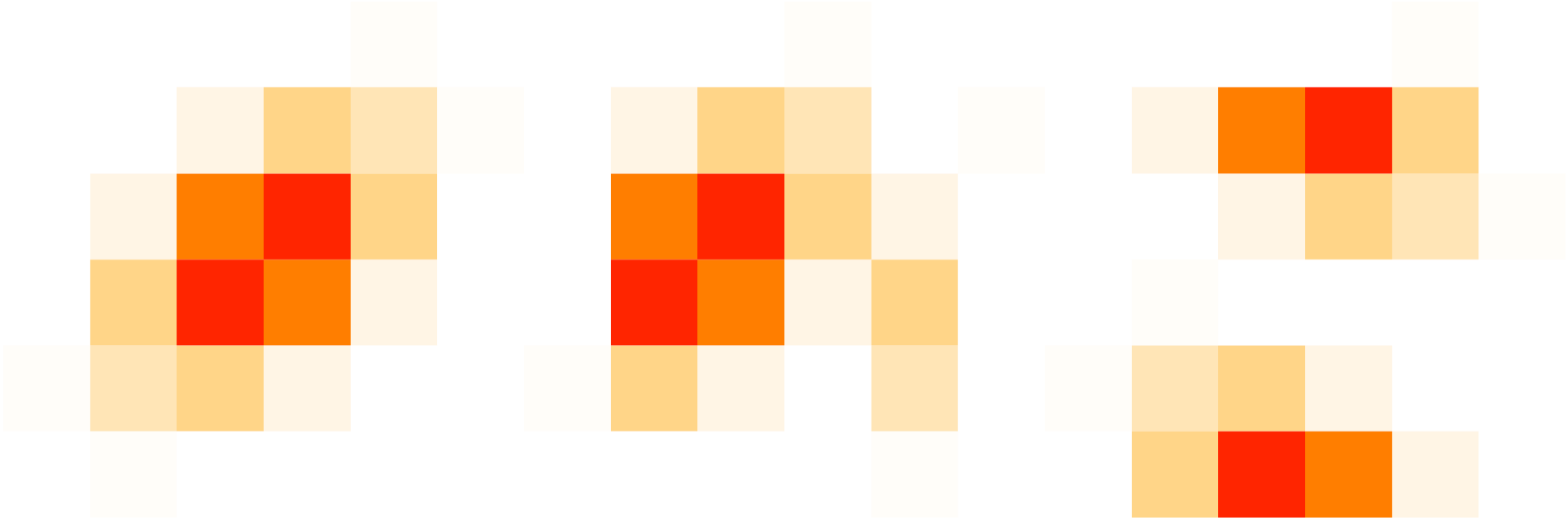
Xによって説明できる、Yの不確実性の割合

- XとYが独立のとき0
- YがXに完全従属するとき1

> 探索的データ解析ソフトウェア**Natto**において、変数間の関係を評価する手法として実装 (Suzuki et al., 2006) https://ef-prime.com/ja/product/

離散化の利点

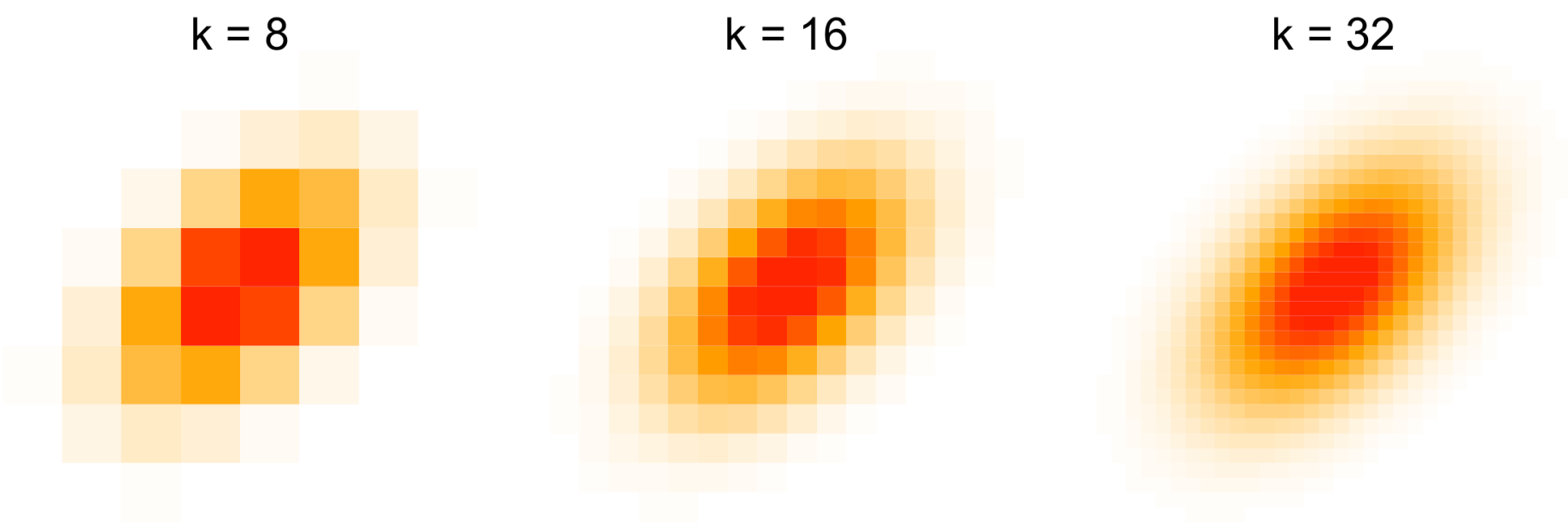
- 変数間の関係を2元分割表で表現
- 行および列の順序を入れ替えても評価は変わらない
  ⇒　非線形な関係を捉えることができる



上記の例は相互情報量や不確実性係数の意味で等価

連続分布の離散近似

- 分割数kを増やせば真の分布のよい近似になる
- 推定時はサンプルサイズnが大きいほど細かく分割できる



相互情報量（連続の場合）

- 非線形な関係も評価することができる
- XとYが独立のとき0
- 完全従属に近づくとき∞に発散

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}\left[\log \frac{f(x,y)}{f_X(x)f_Y(y)}\right] \\
&= \iint f(x,y)\log \frac{f(x,y)}{f_X(x)f_Y(y)}dxdy
\end{aligned}
$$

以下、Iと略記する

相互情報量の離散近似

X, Yの値域をそれぞれk分割した確率変数X_k, Y_kの相互情報量

$$
\begin{aligned}
I(X_k; Y_k) &= \mathbb{E}\left[\log \frac{p(i,j)}{p_X(i)p_Y(j)}\right] \\
&= \sum_{i}^{k}\sum_{j}^{k} p(i,j) \log \frac{p(i,j)}{p_X(i)p_Y(j)}
\end{aligned}
$$

これをI_kと書くと$I_k \leq I$

Iの被積分関数がリーマン可積分なら、分割を細かくすれば$I_k \to I\ (k \to \infty)$

不確実性係数の漸近挙動

不確実係数$U(Y_k|X_k)$について、$k \to \infty$の挙動を考える：

$$U(Y_k|X_k) = \frac{I(X_k;Y_k)}{H(Y_k)}$$

$$0 \leq U(Y_k|X_k) \leq 1$$

- 分子は真の相互情報量$I(X;Y)$に収束し、分母は∞に発散する
- $I(X;Y)$が有限のとき、不確実性係数は0に収束してしまう

> MIC (Reshef et al., 2011) も変数を離散化し、相互情報量を基準化する。基準化した値の最大値をとることで、分母の発散に対応していると解釈できる
>
> $$\max_{s,t} \frac{\hat{I}\left(X_s;Y_t\right)}{\log(\min\{s,t\})}$$

方針の再検討

理論値としての評価指標を定め、その推定手法を考える

- 確率変数の型を問わない評価指標を定義
  - 同時分布の汎関数によって定められる量
- 離散近似は推定手法の候補として位置付ける
  - サンプルサイズnに応じて分割数kを大きくできると想定
  - $k \to \infty$の極限が意味のある値になるように指標を設計

相互情報量（一般の場合）

$$I(X;Y) = \begin{cases} \mathbb{E}_{P_{X,Y}}\left[\log \dfrac{dP_{X,Y}}{d(P_X \otimes P_Y)}\right] & (P_{X,Y} \ll P_X \otimes P_Y) \\ \infty & (\text{otherwise}) \end{cases}$$

dP/dQはラドン＝ニコディム微分

- $I \in [0, \infty]$
- $I = 0 \Leftrightarrow X$とYは独立、関係が強いとき大きな値

> 対数変換の扱いに工夫を要する場合がある。例として推定量の分布を評価する際に多項式近似が用いられる (Hamdan & Tsokos, 1971)

提案：相互依存度（Mutual Dependency）

$$\psi(X;Y) := \begin{cases} \mathbb{E}_{P_{X,Y}}\left[\dfrac{dP_{X,Y}}{d(P_X \otimes P_Y)}\right] & (P_{X,Y} \ll P_X \otimes P_Y) \\ \infty & (\text{otherwise}) \end{cases}$$

相互情報量から対数変換を除いたものとして定義

- $\psi \in [1, \infty]$
- $\psi = 1 \Leftrightarrow X$と$Y$は独立、関係が強いとき大きな値

逆数は$\psi^{-1} \in [0, 1]$で、独立性の指標となり得る
対数変換を含まないことで、式展開や推定量の評価が容易になる

相互依存度の解釈

- 連続の場合：仮説「XとYが独立」に対する密度比の期待値
- 離散の場合：相関ルールにおけるリフトの期待値

$$\psi(X;Y) = \mathbb{E}_{P_{X,Y}}\left[\frac{f(x,y)}{f_X(x)f_Y(y)}\right] = \mathbb{E}_{P_{X,Y}}\left[\frac{f_{Y|X}(y|x)}{f_Y(y)}\right]$$

Jensenの不等式による相互情報量の上界を構成する：

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{P_{X,Y}}\left[\log \frac{dP_{X,Y}}{d(P_X \otimes P_Y)}\right] \\ &\leq \log\left(\mathbb{E}_{P_{X,Y}}\left[\frac{dP_{X,Y}}{d(P_X \otimes P_Y)}\right]\right) \\ &= \log \psi(X;Y) \end{aligned}$$

f-ダイバージェンスによる整理

- 相互情報量はKLダイバージェンスによって定義される
- 相互依存度はχ^2ダイバージェンスに定数1を足した値

$$I(X;Y) = D_{\mathrm{KL}}(P_{X,Y} \| P_X \otimes P_Y)$$
$$\psi(X;Y) = D_{\chi^2}\left(P_{X,Y} \| P_X \otimes P_Y\right) + 1$$

$\Rightarrow$ f-ダイバージェンスやχ^2統計量の性質が利用できる

相互依存度の性質

相互情報量と同様の望ましい性質をもつ：

- 確率変数の型を問わず定義される（連続、離散あるいは混合）
- 変数変換に対して不変（全単射をなす可測変換）
- 非線形な関係も評価することができる

> 相互情報量を用いても類似の議論が可能だが、
> 対数変換を含まないことで各種の議論がシンプルになる

相互依存度（連続の場合）

(X, Y)が同時密度$f(x, y)$、周辺密度$f_X(x), f_Y(y)$をもつとき

$$\begin{aligned}\psi(X;Y) &= \mathbb{E}_{P_{X,Y}}\left[\frac{f(x,y)}{f_X(x)f_Y(y)}\right] \\ &= \iint \frac{f(x,y)^2}{f_X(x)f_Y(y)}dxdy\end{aligned}$$

(X, Y)が相関係数ρをもつ2変量正規分布に従うとき

$$\psi = \frac{1}{1-\rho^2}$$

が成り立ち、ψとρ^2は一対一に対応する。すなわち

$$\rho^2 = 1 - \psi^{-1}$$

検討：相互依存度に基づく相関指標

正規分布における議論に基づいて以下を定義（＊）：

$$r_\psi = \sqrt{1 - \psi^{-1}}$$

- (X, Y)が2変量正規分布に従うとき、相関係数の絶対値$|\rho|$と一致
- $[0, 1]$に値をとり、XとYが独立のとき0
- XとYがともに無限集合に値をとり、互いに完全従属するとき1
  - 有限集合に値をとる場合、完全従属であっても$r_\psi < 1$

> ＊ 相互情報量に基づく Linfoot (1957) の情報相関係数（informational coefficient of correlation）と同様の発想
>
> $$r_I = \sqrt{1 - e^{-2I(X;Y)}}$$

相互依存度（離散の場合）

(X, Y)が同時確率$p(x, y)$、周辺確率$p_X(x), p_Y(y)$をもつとき

$$\begin{aligned}\psi(X;Y) &= \mathbb{E}_{P_{X,Y}}\left[\frac{p(x,y)}{p_X(x)p_Y(y)}\right] \\ &= \sum_x \sum_y \frac{p(x,y)^2}{p_X(x)p_Y(y)}\end{aligned}$$

X, Yの支持集合の基数をそれぞれk_X, k_Yとするとき

$$1 \leq \psi(X;Y) \leq \min\{k_X, k_Y\}$$

YがXに完全従属する、すなわち$p_{Y|X}$が一点分布となるとき

$$\psi(X;Y) = k_Y \leq k_X$$

> 一方が連続または無限離散などの場合も同様の議論が可能

提案：予測可能性の指標

相互依存度に基づく**予測スコア（Predictability Score）**を以下で定義する：

$$\mathrm{Pred}_{\psi}(X \to Y) := \begin{cases} \sqrt{\dfrac{1 - \psi^{-1}}{1 - k_Y^{-1}}} & (k_Y > 1) \\ 1 & (k_Y = 1) \end{cases}$$

k_YはYの支持集合の基数（無限集合の場合は∞）

- $[0, 1]$に値をとる
- XとYが独立のとき0、YがXに完全従属するとき1
- (X, Y)が2変量正規分布に従うとき、相関係数の絶対値$|\rho|$と一致

対称性をもつ指標

予測スコアは予測可能性の非対称性が反映されているが、
相関係数のように対称性をもつ指標があると便利

以下を満たす指標を**一般化相関尺度（Generalized Correlation Measures）**
と呼ぶことにする：

- $[0, 1]$に値をとる
- XとYが独立のとき0、***互いに***完全従属するとき1
- (X, Y)が2変量正規分布に従うとき、相関係数の絶対値$|\rho|$と一致

提案：相互依存度に基づく一般化相関尺度

相互依存度に基づく**一般化相関尺度**を以下で定義する：

$$\mathrm{gCor}_{\psi}(X,Y) := \begin{cases} \sqrt{\dfrac{1-\psi^{-1}}{1-\left(\sqrt{k_X k_Y}\right)^{-1}}} & (k_X k_Y > 1) \\ 1 & (k_X k_Y = 1) \end{cases}$$

k_X, k_YはX, Yの支持集合の基数（無限集合の場合は∞）

- $[0,1]$に値をとる
- XとYが独立のとき0、互いに完全従属するとき1
- (X,Y)が2変量正規分布に従うとき、相関係数の絶対値$|\rho|$と一致

以降$\mathrm{gCor}(X,Y)$または単にgCorと略記

相互依存度の離散近似

連続確率変数X, Yの値域をそれぞれk分割したX_k, Y_kの相互依存度

$$\psi(X_k; Y_k) = \mathbb{E}\left[\frac{p(i,j)}{p_X(i)p_Y(j)}\right]$$
$$= \sum_i^k \sum_j^k \frac{p(i,j)^2}{p_X(i)p_Y(j)}$$

これをψ_kと書くと$\psi_k \leq \psi$

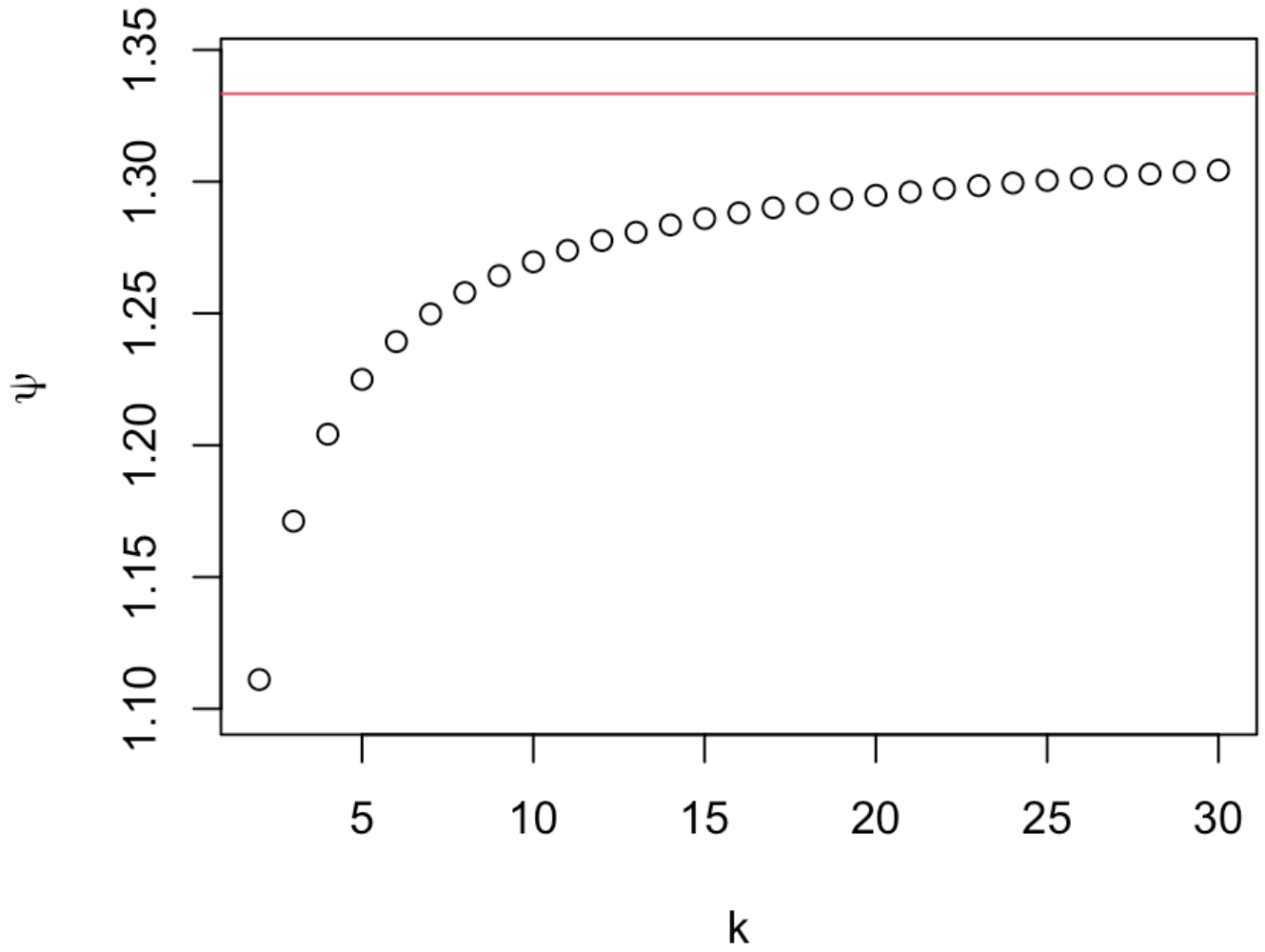
ψの被積分関数がリーマン可積分なら$\psi_k \to \psi \ (k \to \infty)$

一般化相関尺度（および予測スコア）についても同様：

$$\mathrm{gCor}(X_k, Y_k) \to \mathrm{gCor}(X, Y) \ \ (k \to \infty)$$

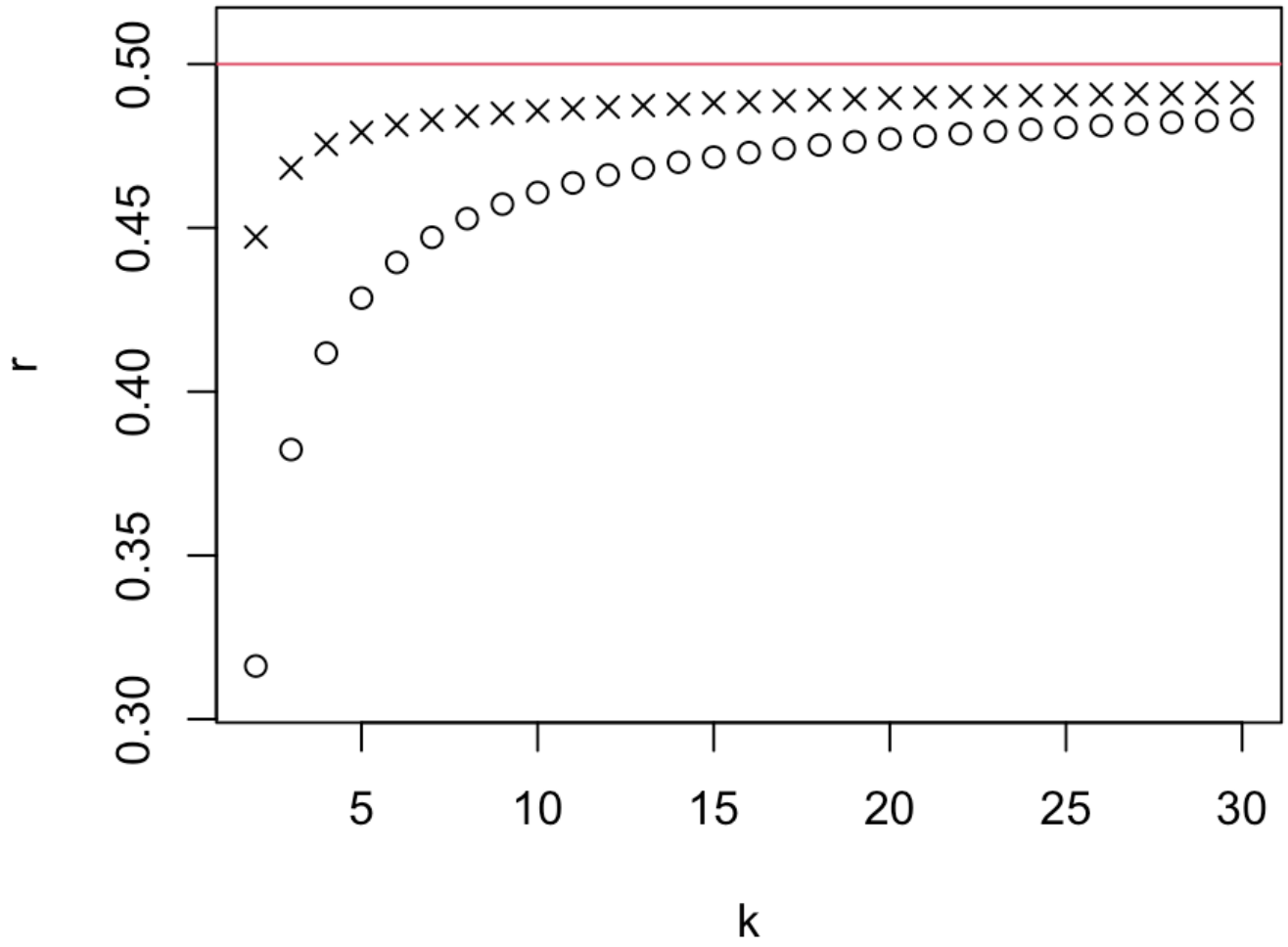
数値例：相互依存度の近似

相関係数$\rho = 0.5$をもつ2変量正規分布をk分位点で分割
○が離散化による近似値ψ_k、赤線が真値$\psi = 4/3$



数値例：一般化相関尺度の近似

○は離散化による近似値ψ_kを式$\sqrt{1-\psi^{-1}}$に代入した値
×が$\sqrt{1-k^{-1}}$で基準化したgCorの値、赤線が真値$|\rho| = 0.5$



離散近似からの示唆：指標の解釈

離散確率変数V, Wが（観測されない）連続確率変数X, Yの分割

$$V = v(X),\ W = w(Y)$$

と仮定すれば、$\mathrm{gCor}(V, W)$を$\mathrm{gCor}(X, Y)$の近似として解釈できる

> 特に(X, Y)が2変量正規分布に従うとき、相関係数の絶対値$|\rho|$に対する近似となる

$\Rightarrow$連続、離散および混合型に対する統一的な評価に意味を与える

> 例：$\mathrm{gCor}(V, W) = 0.7$のとき、背後に$\mathrm{Cor}(X, Y) = 0.7$となる正規確率変数$X, Y$がある状況に近いと考える

離散近似からの示唆：離散化による推定

数値データ（実数または整数）を分割して離散化

離散化データに対してgCorなどの指標を推定

プラグイン推定量$\hat{\psi}$はχ^2統計量で表すことができる：

$$\hat{\psi} = \sum_i \sum_j \frac{n_{ij}^2}{n_{i.} n_{.j}} = \frac{\chi^2}{n} + 1$$

$\Rightarrow$　近似的にχ^2分布を用いた統計的推測が可能

$$n_{ij} = \#(X = x_i, Y = y_j),\ \ n_{i.} = \#(X = x_i),\ \ n_{.j} = \#(Y = y_j)$$

$$n = \sum_i \sum_j n_{ij},\ \ \chi^2 = \sum_i \sum_j \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}/n}$$

連関係数との関係

X, Yが離散のとき、クラメールのVはχ^2統計量で表される：

$$V = \sqrt{\frac{\chi^2/n}{\min\{k_X, k_Y\} - 1}}$$

$k_X = k_Y = 2$のときピアソンのϕが定義され、$V = |\phi|$となる

相互依存度のプラグイン推定量はVの分子に1を足した値となり、
逆数をとって変換することで一般化相関尺度の推定値が得られる：

$$\hat{\psi} = \frac{\chi^2}{n} + 1$$

$$\widehat{\mathrm{gCor}}(X, Y) = \sqrt{\frac{1 - \hat{\psi}^{-1}}{1 - \left(\sqrt{k_X k_Y}\right)^{-1}}}$$

相関係数との関係

(X, Y)が相関係数ρをもつ2変量正規分布に従うとき、k分割による離散化を行えば、適当な条件のもとで以下が得られる：

$$\sqrt{1 - \hat{\psi}_k^{-1}} = \sqrt{1 - \left(\frac{\chi^2}{n} + 1\right)^{-1}}$$

$$= \sqrt{\frac{\chi^2/n}{1 + \chi^2/n}}$$

$$\to |\rho| \quad (k, n \to \infty)$$

> K. Pearson (1904) において$\phi^2 = \chi^2/n$をmean square contingencyと定義し、類似の議論を展開している。上記に相当する値をfirst coefficient of contingencyとして提案しており、$\widehat{\mathrm{gCor}}$はこの値を分割数で基準化したものにあたる

推定手法の提案

分位点グリッド近似 (Quantile Grid Approximation)

1. 分割数kを定め、数値データを標本k分位点で分割する
2. 分割データについて相互依存度ψを推定する
3. ψの推定値を用いて一般化相関尺度などを算出する

- $k, n \to \infty$のもとで一致性をもつ
- kの値はサンプルサイズnに応じて決める
  - 例：分割表のセルあたり平均サンプルサイズを50以上に保つ

$$\max\left\{2, \left\lfloor\sqrt{n/50}\right\rfloor\right\}$$

推定シミュレーション

以下を1,000回繰り返す：

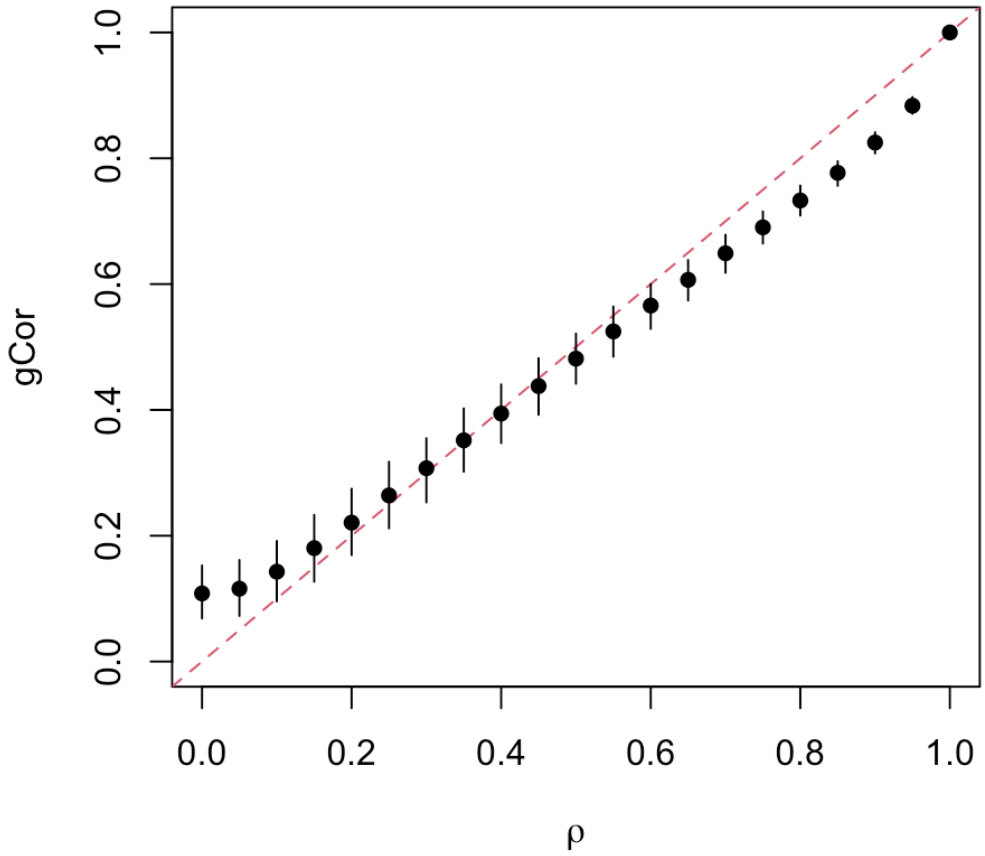
- 平均0、分散1の2変量正規分布に相関係数ρを0.05刻みで設定
- サンプルサイズnのデータを抽出
- 分割数kを定め、$\widehat{\mathrm{gCor}}(X_k, Y_k)$を計算

$\Rightarrow$　推定値の平均値、および90%区間をプロットして確認

以降、セルあたり平均サンプルサイズを**平均サイズ**と記載

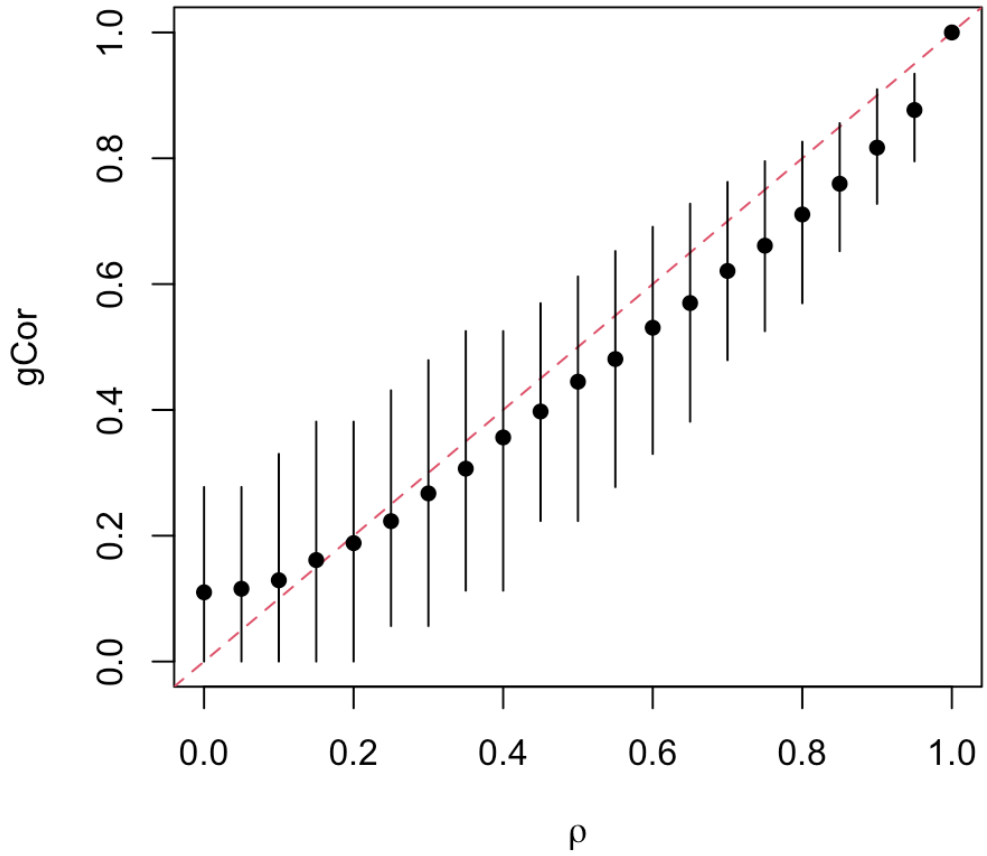
n = 1,000, k = 4

$\sqrt{1000/50} \approx 4.47$を参考に$k$を設定、平均サイズ62.5
赤の点線は真値。バイアスはあるものの真値に近く、バラツキも小さい



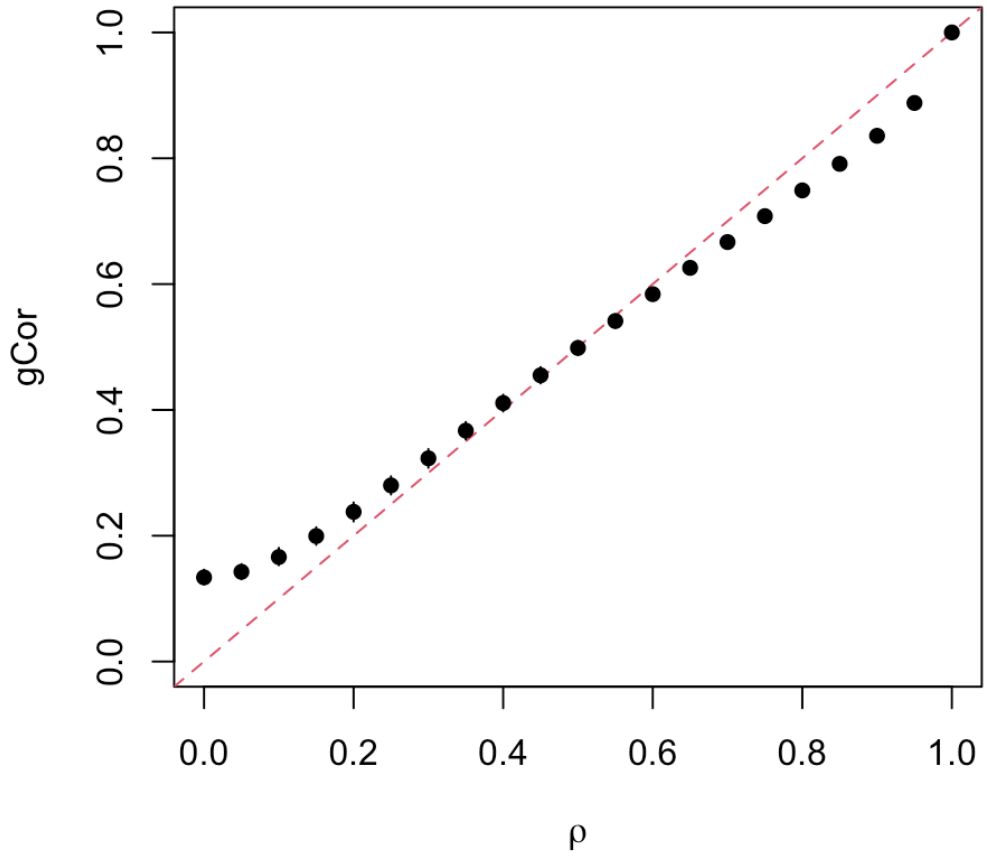
n = 100, k = 2

$\sqrt{100/50} \approx 1.41$を参考に$k$を設定、平均サイズ25
バラツキが大きくなるものの、バイアスはそこまで増加しない



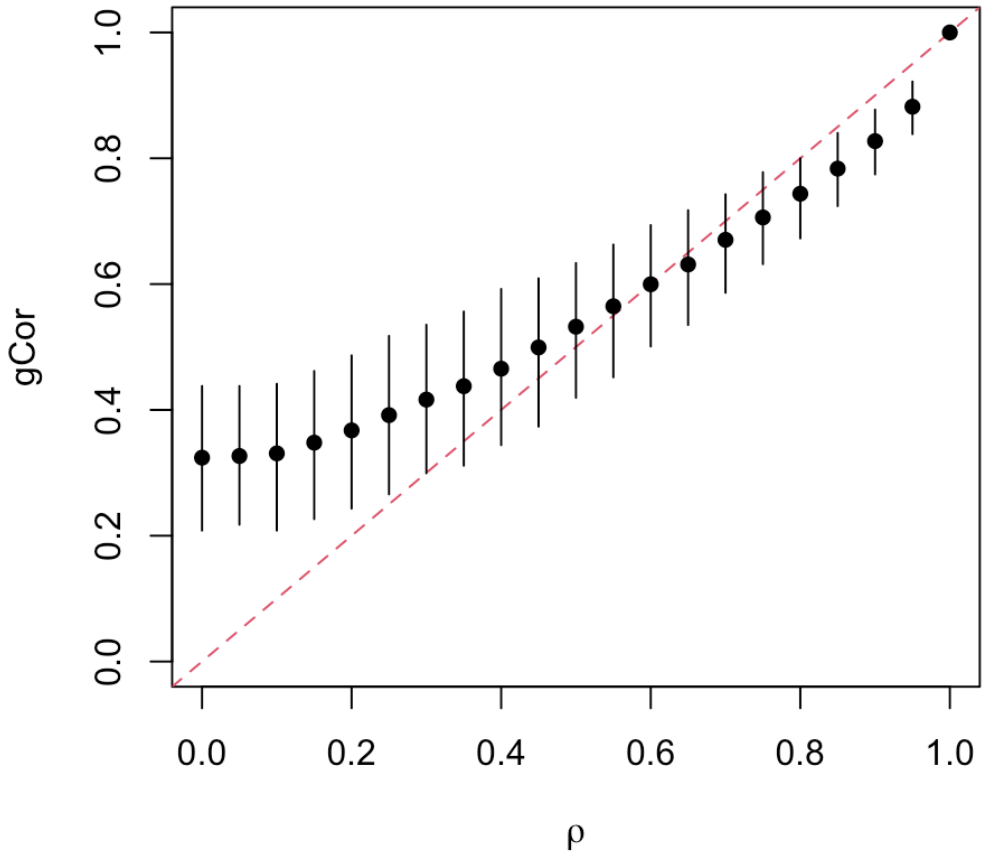
n = 10,000, k = 14

$\sqrt{10000/50} \approx 14.14$を参考に$k$を設定、平均サイズ約51
バイアスは残るが、バラツキはほとんどなくなる



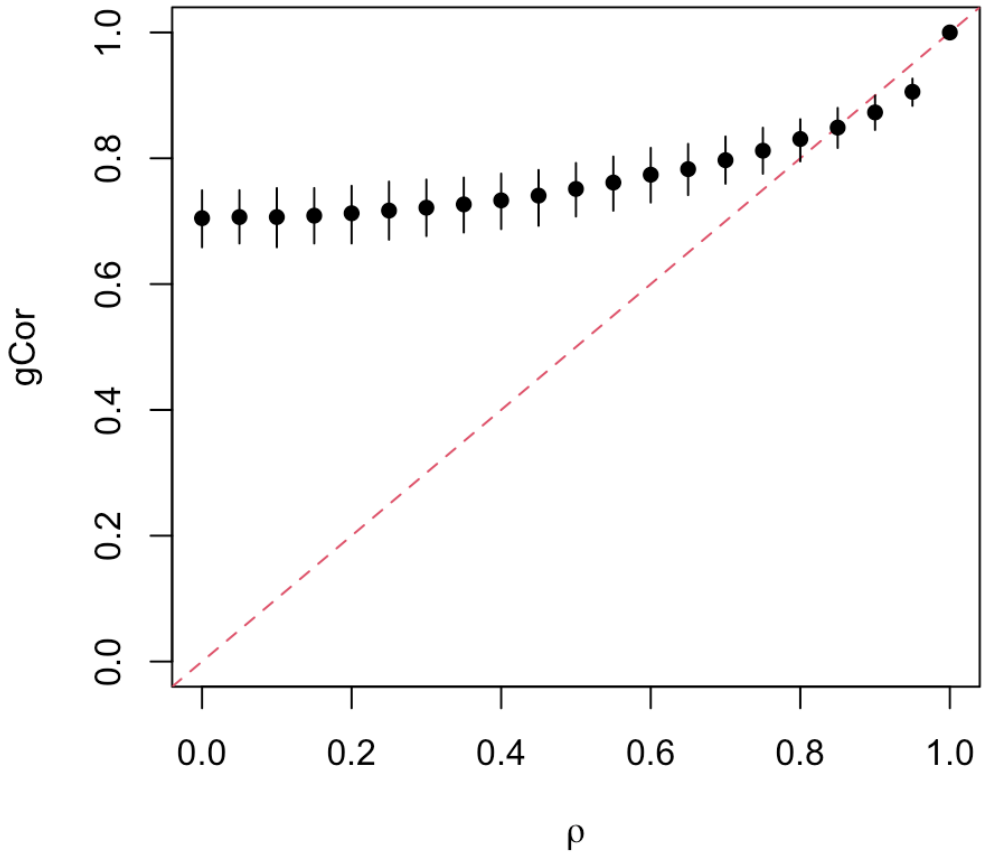
n = 100, k = 4

$\sqrt{100/50} \approx 1.41$を大幅に上回り、平均サイズは6.25
バイアスは大きくなるが、平均的な順序関係は保たれている



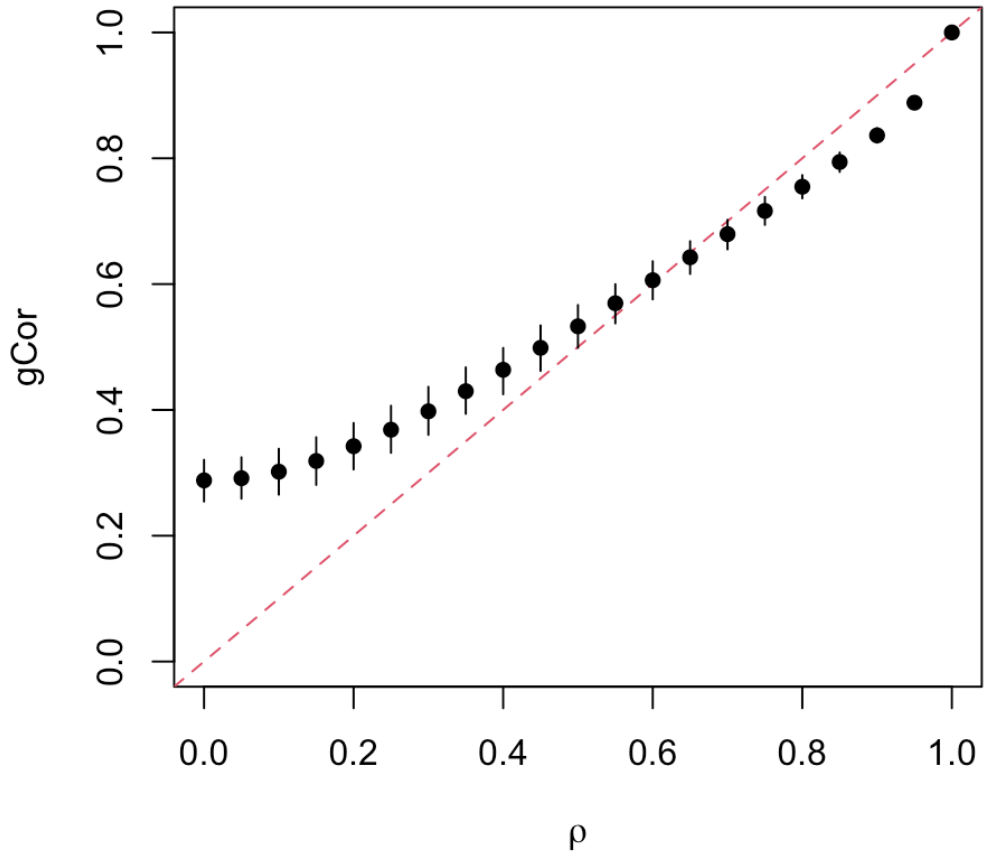
n = 100, k = 10

平均サイズが1となる極端な設定
バイアスが非常に大きくなり、バラツキはかえって小さくなる



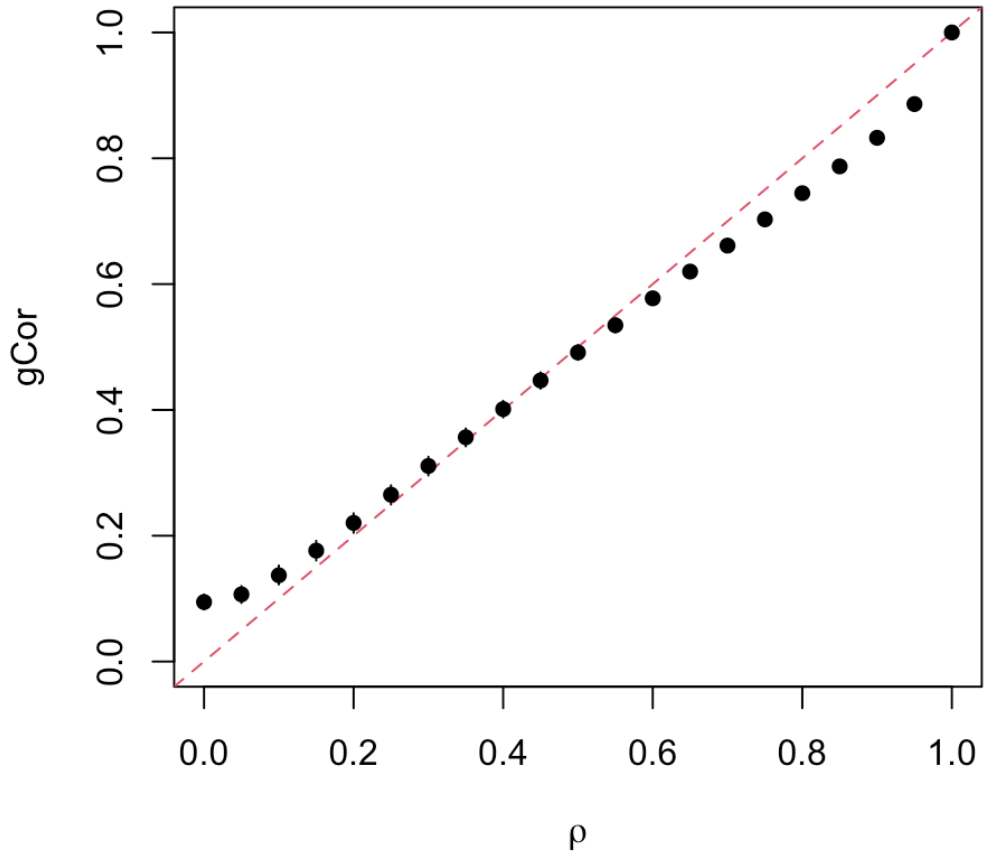
n = 1,000, k = 10

$\sqrt{1000/50} \approx 4.47$に対して約2.2倍に$k$を設定、平均サイズ10
バイアスの傾向はn = 100, k = 4の例（平均サイズ6.25）と類似



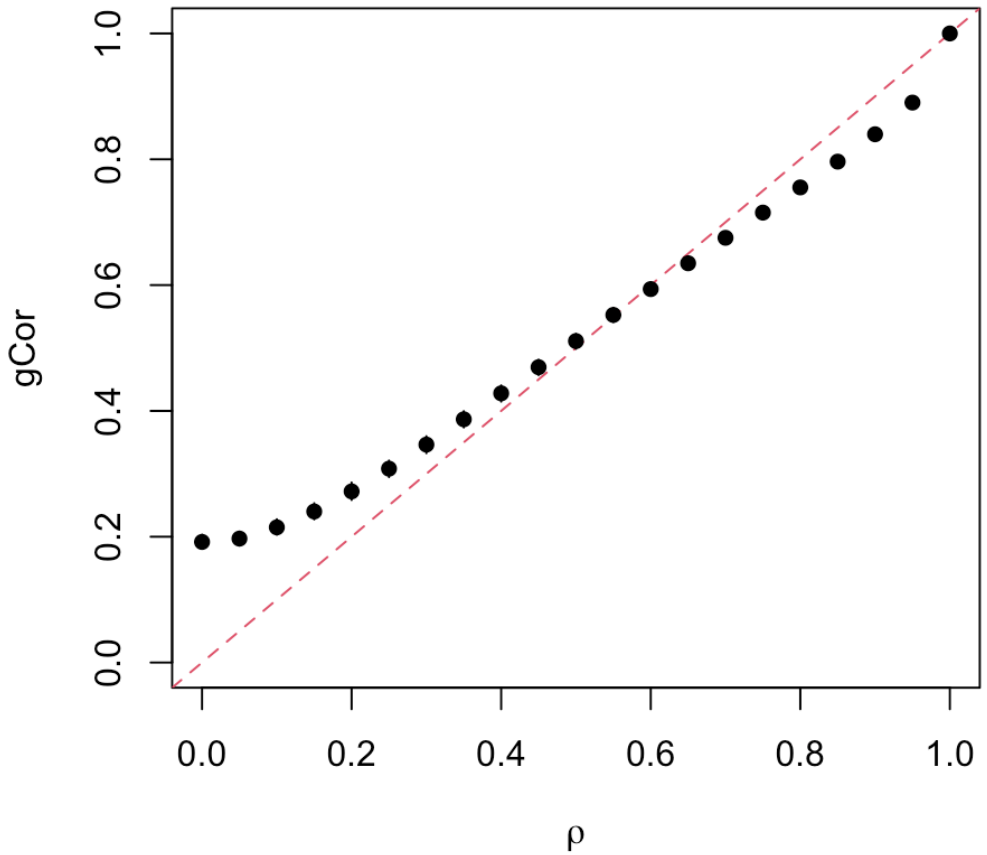
n = 10,000, k = 10

nを10倍に増やし、$\sqrt{10000/50} \approx 14.14$に対して約0.7倍に$k$を設定
平均サイズは100に増加し、バイアスが大きく改善



n = 10,000, k = 20

kを2倍に増加し、平均サイズは25に減少
バイアスが増加するが、非線形な関係を検知しやすくなると期待



推定シミュレーションのまとめ

- サンプルサイズnと分割数kによってバイアスが変動
- セルあたりの平均サンプルサイズが50〜100程度のとき良好
- バイアスが大きい場合でも、順序関係は保たれる傾向

非線形な関係を検出するためには分割を増やす必要がある
バイアスを許容して分割を増やし、関係の検知性能を重視する方針も

- 例：データ解析の初期段階における探索的解析
  - 関係がありそうな変数の組を列挙、可視化やモデリングに繋げる

数値例

ランダム生成した2次元の数値データを評価

サンプルサイズn、分割数kを動かして挙動を確認

標本相関係数と比較

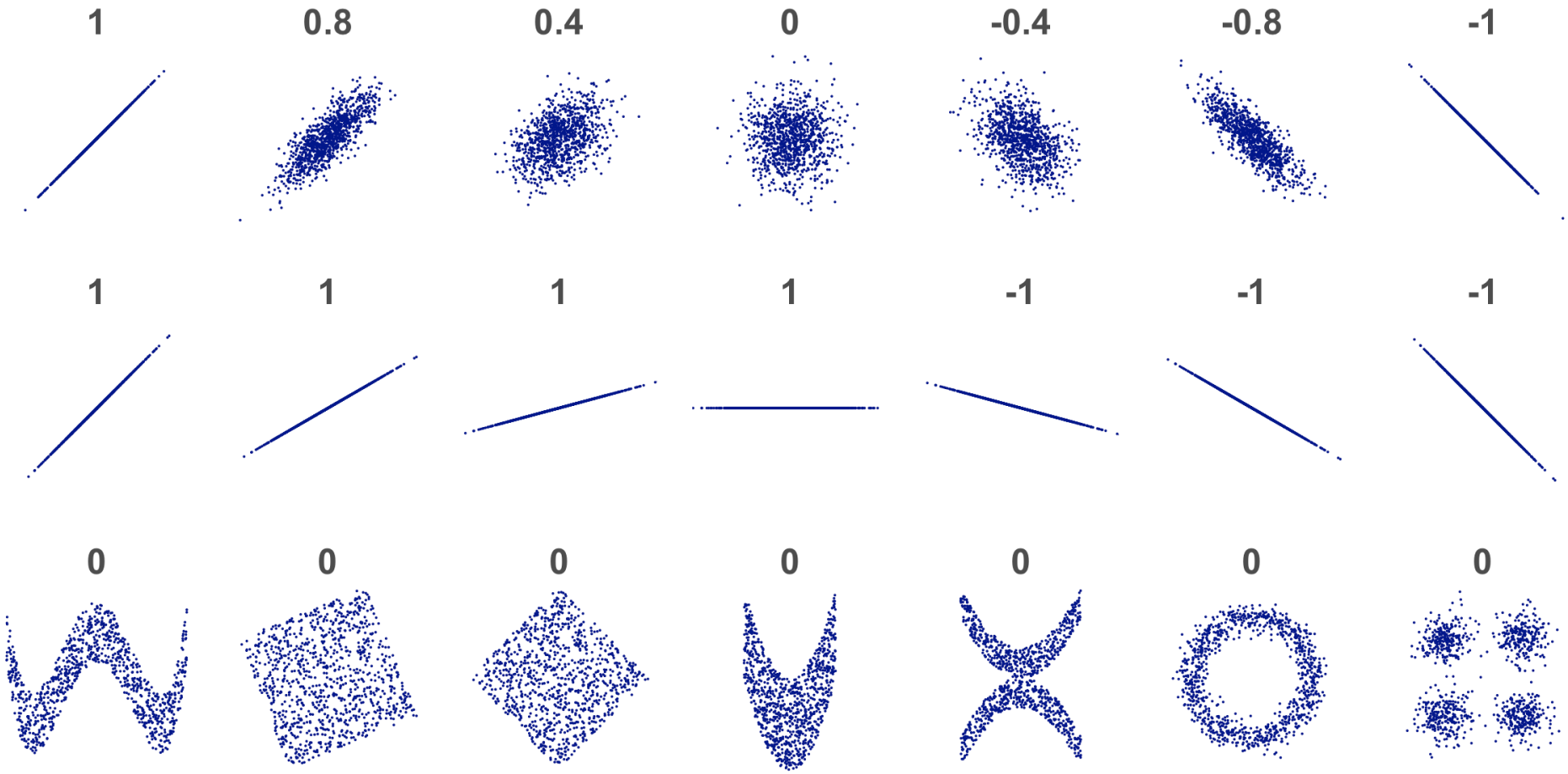
Wikipedia「ピアソンの積率相関係数」掲載の例をもとに作成

- 出典：https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg
  - CC0 1.0 Universal（パブリックドメイン）

n = 1,000　標本相関係数

上段は相関係数ρの絶対値を1, 0.8, 0.4, 0に設定した2変量正規分布

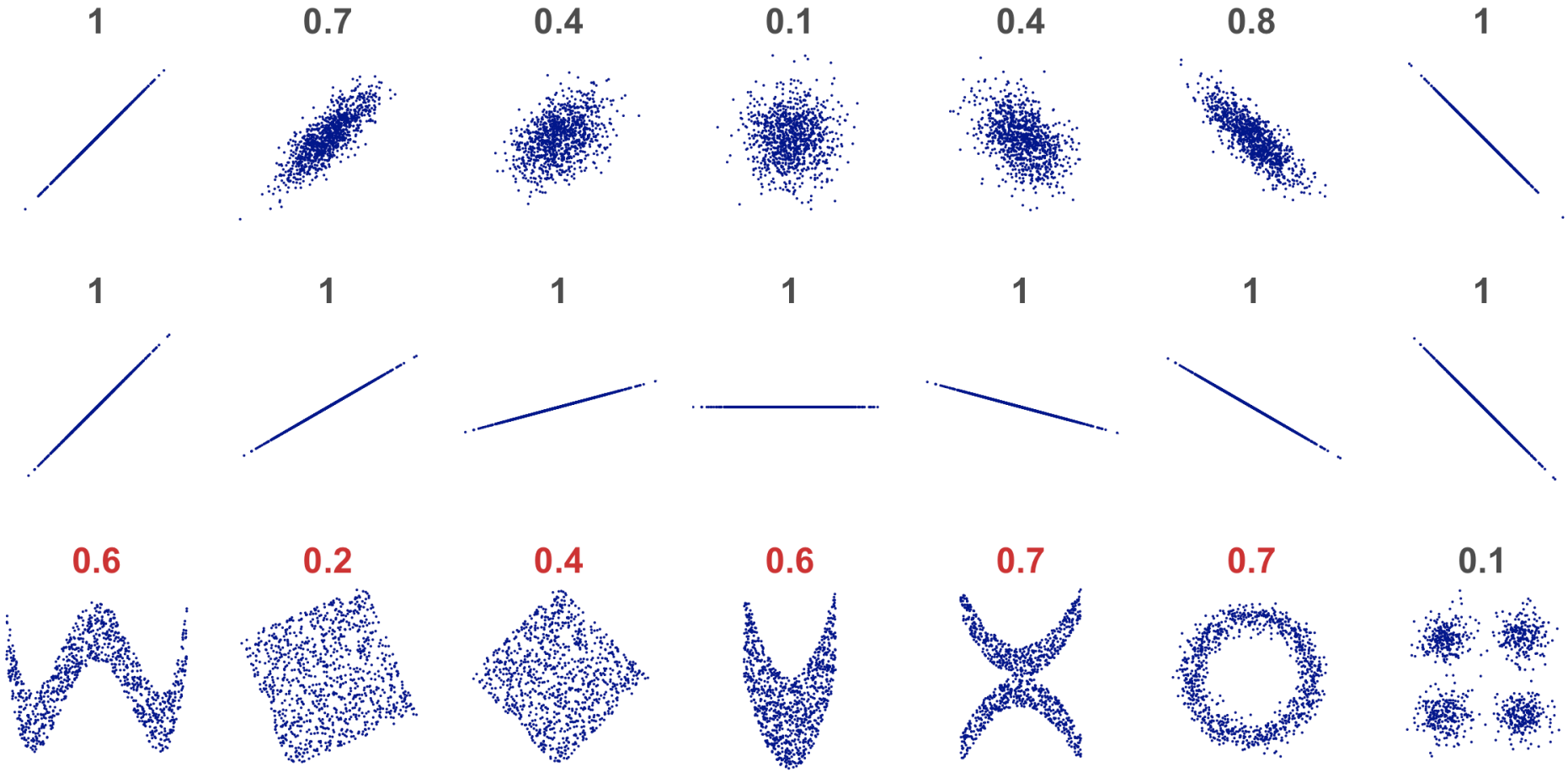
小数点以下は1桁に丸めて表示（以降も同様）



n = 1,000, k = 4　一般化相関尺度

平均サイズ50を基準としてkを設定（$n/k^2 = 62.5$）

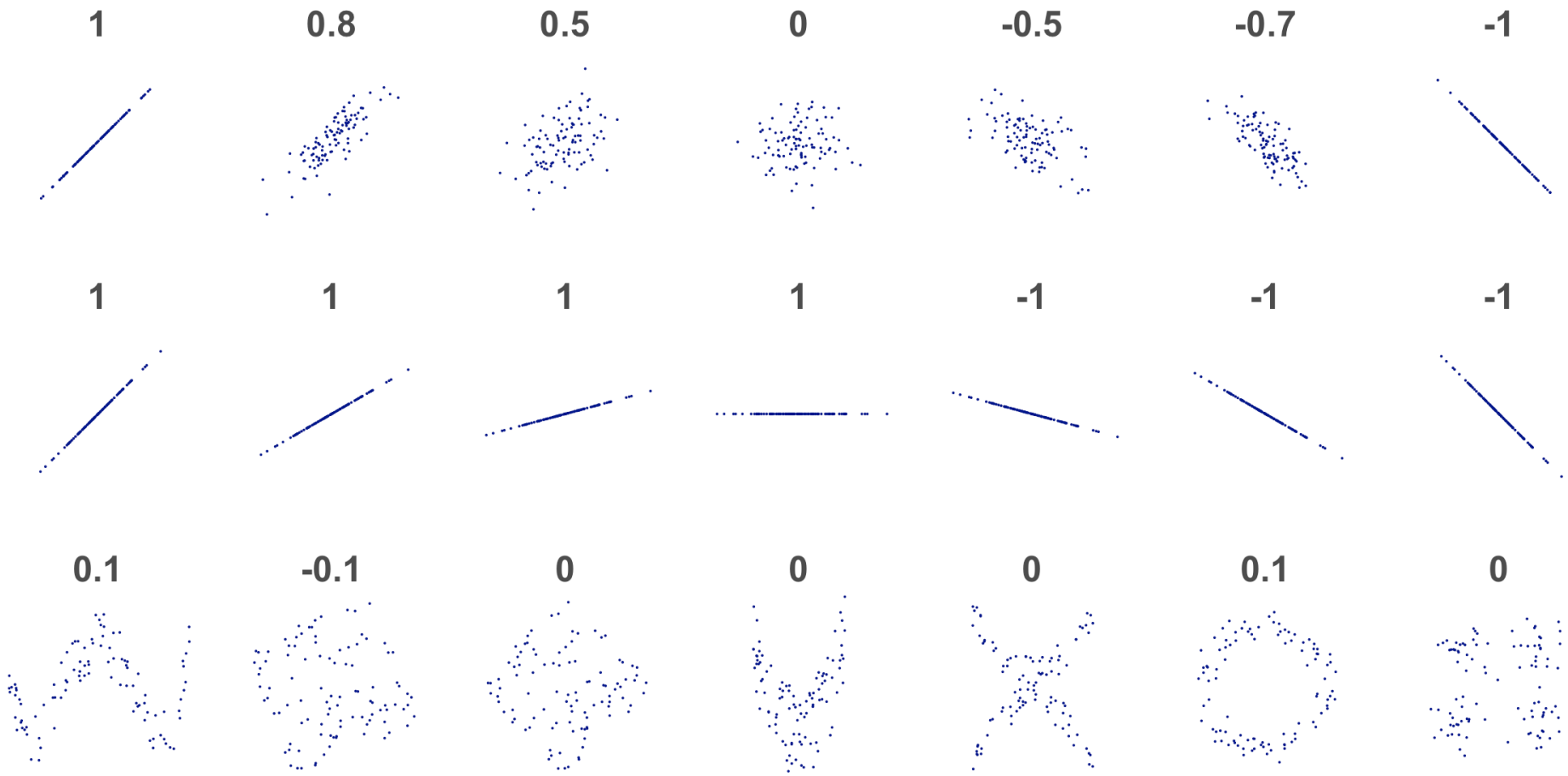
非線形なパターンを検出しつつ、正規分布に対しては$|\rho|$に近い値



n = 100　標本相関係数

サンプルサイズが小さい場合

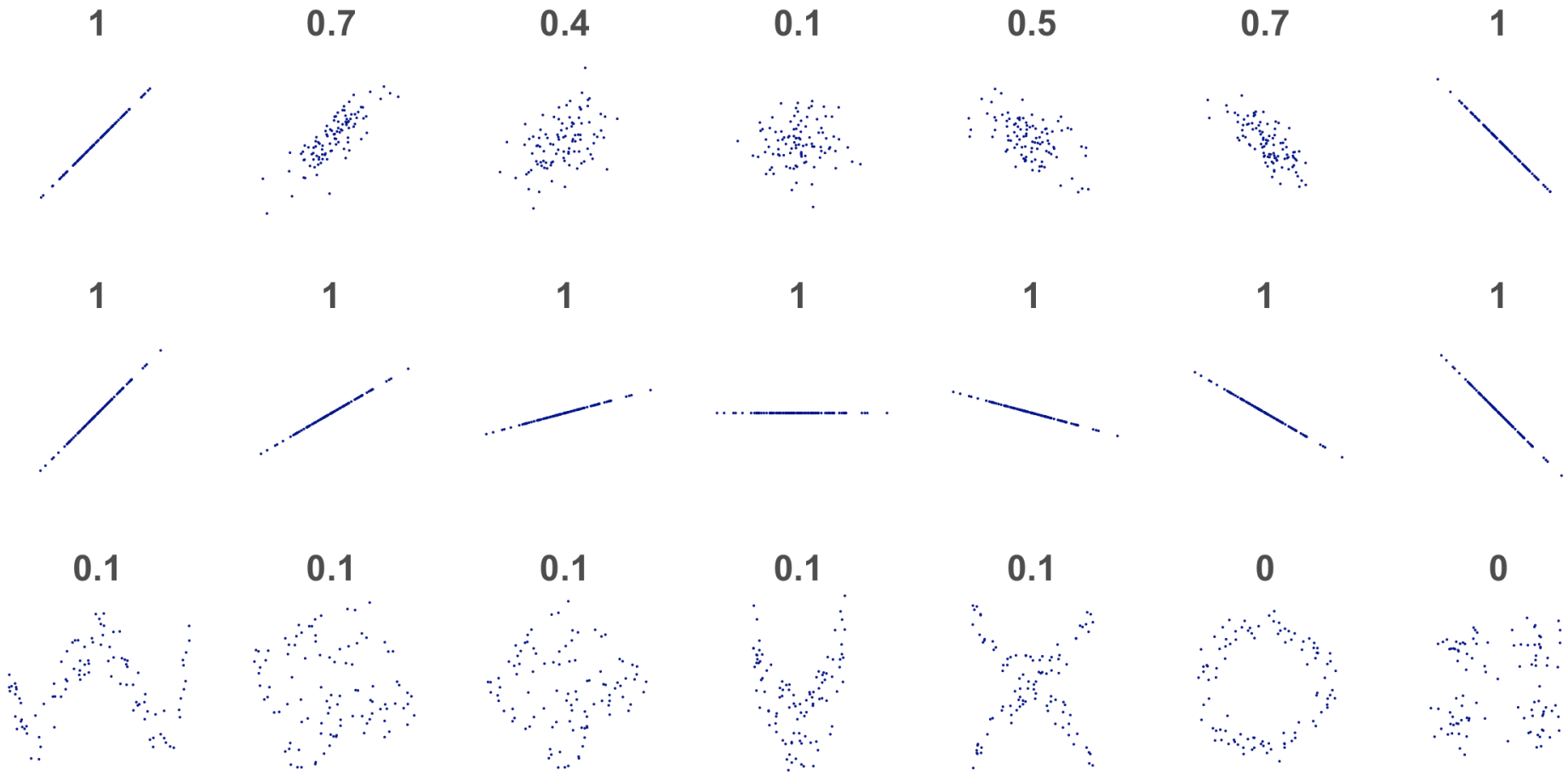
正規乱数に対しても多少の推定誤差が生じる



n = 100, k = 2　一般化相関尺度

平均サイズ50を基準としてkを設定（$n/k^2 = 25$）

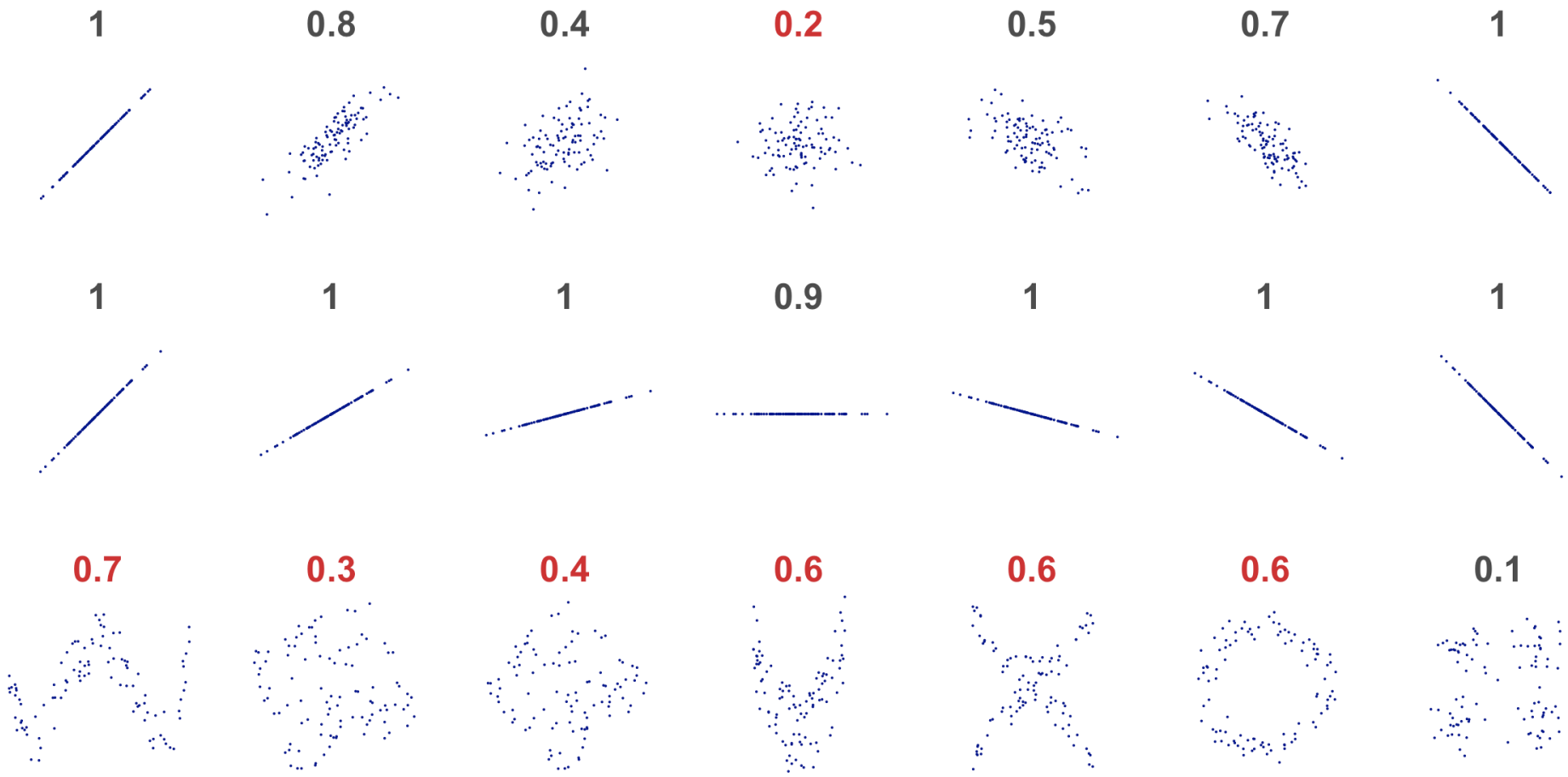
正規分布に対しては$|\rho|$に近いが、非線形パターンを検出できていない



n = 100, k = 3　一般化相関尺度

平均サイズ50を大幅に下回る（$n/k^2 \approx 11$）

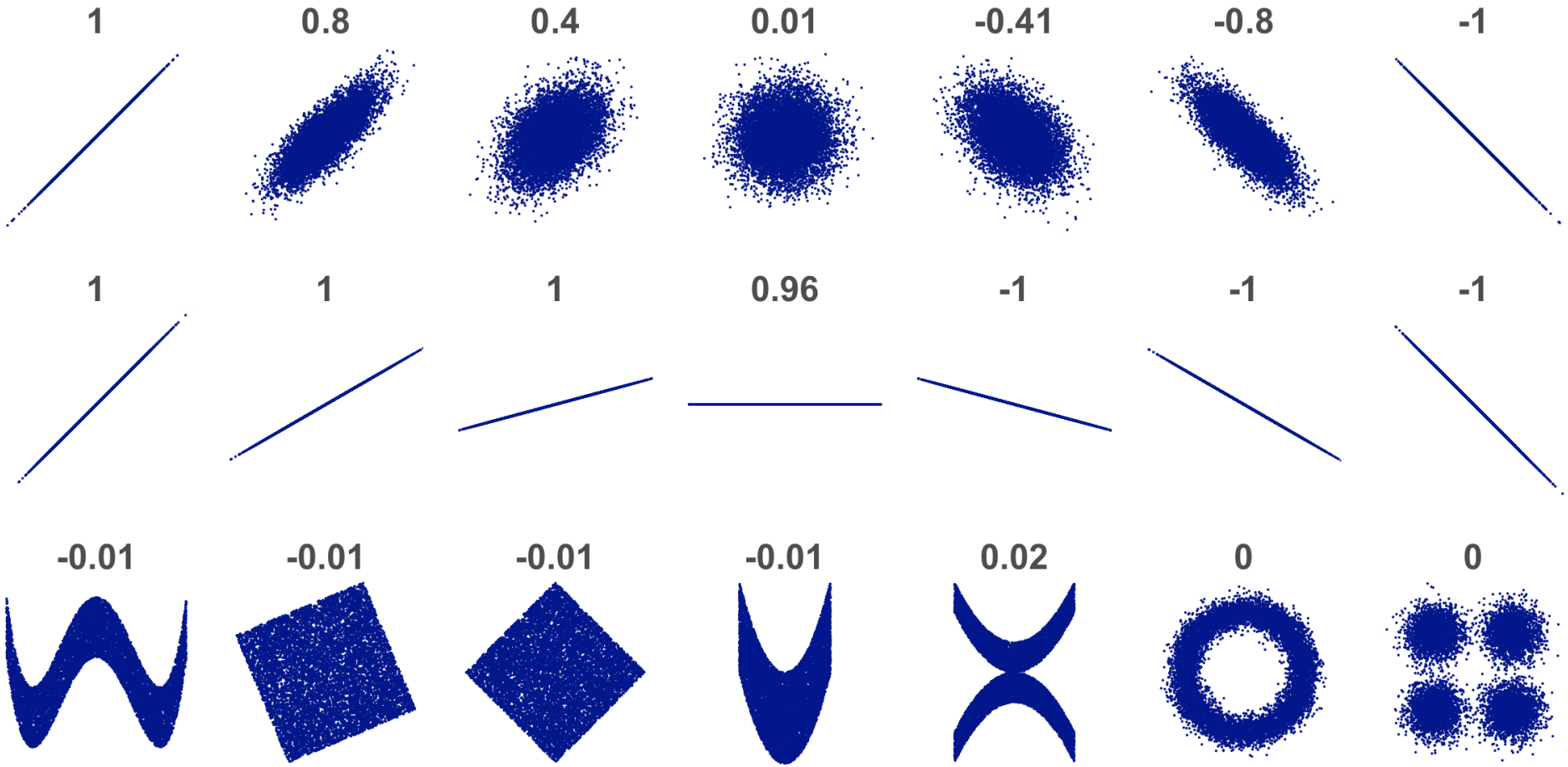
近似バイアスが増加するものの、非線形パターンの検出には成功



n = 10,000　標本相関係数

サンプルサイズが大きい場合について、小数点以下第2位まで表示

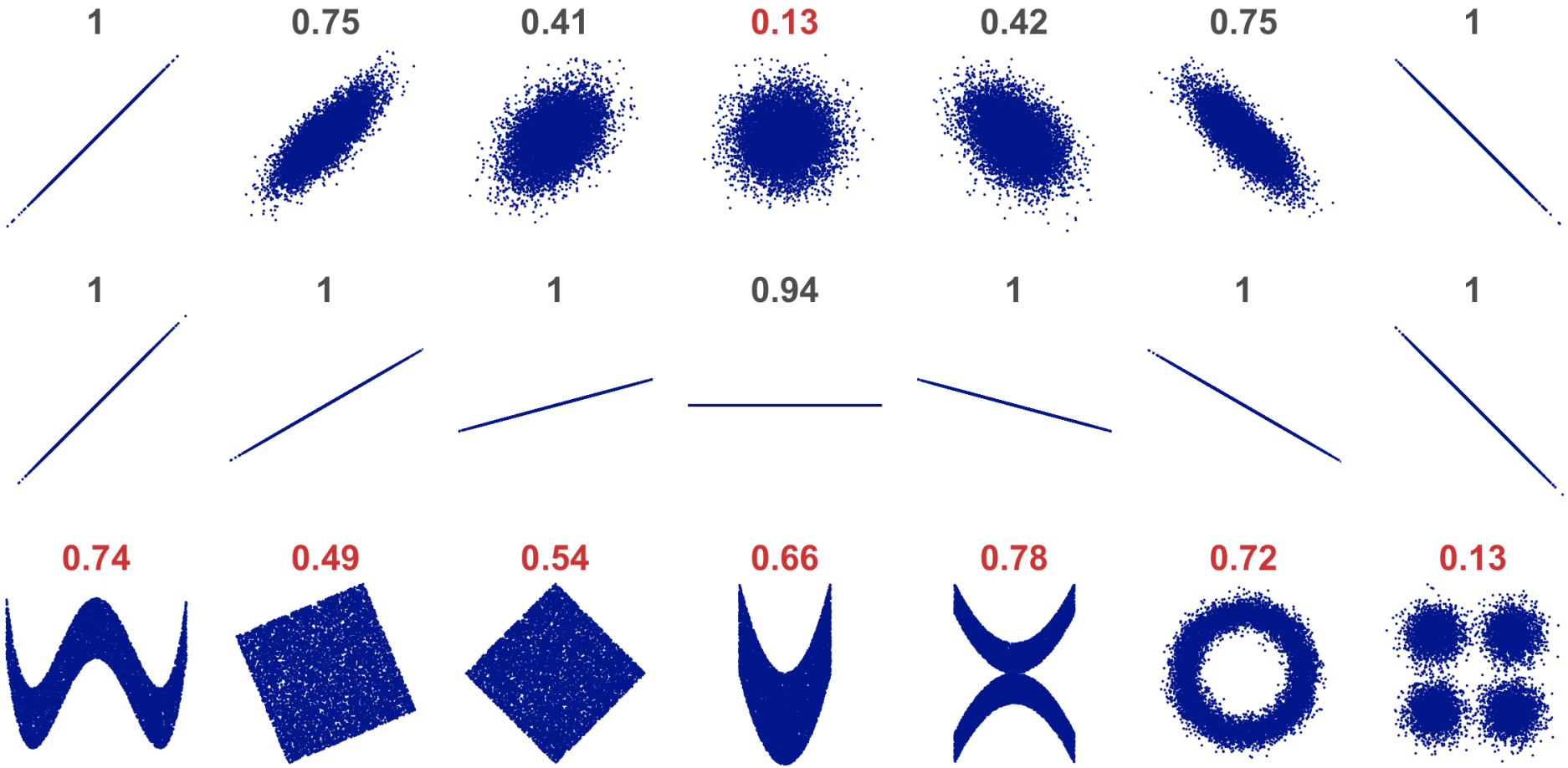
サンプルサイズを増やしても、非線形パターンについては0に近い



n = 10,000, k = 14　一般化相関尺度

平均サイズ50を基準としてkを設定（$n/k^2 \approx 51$）

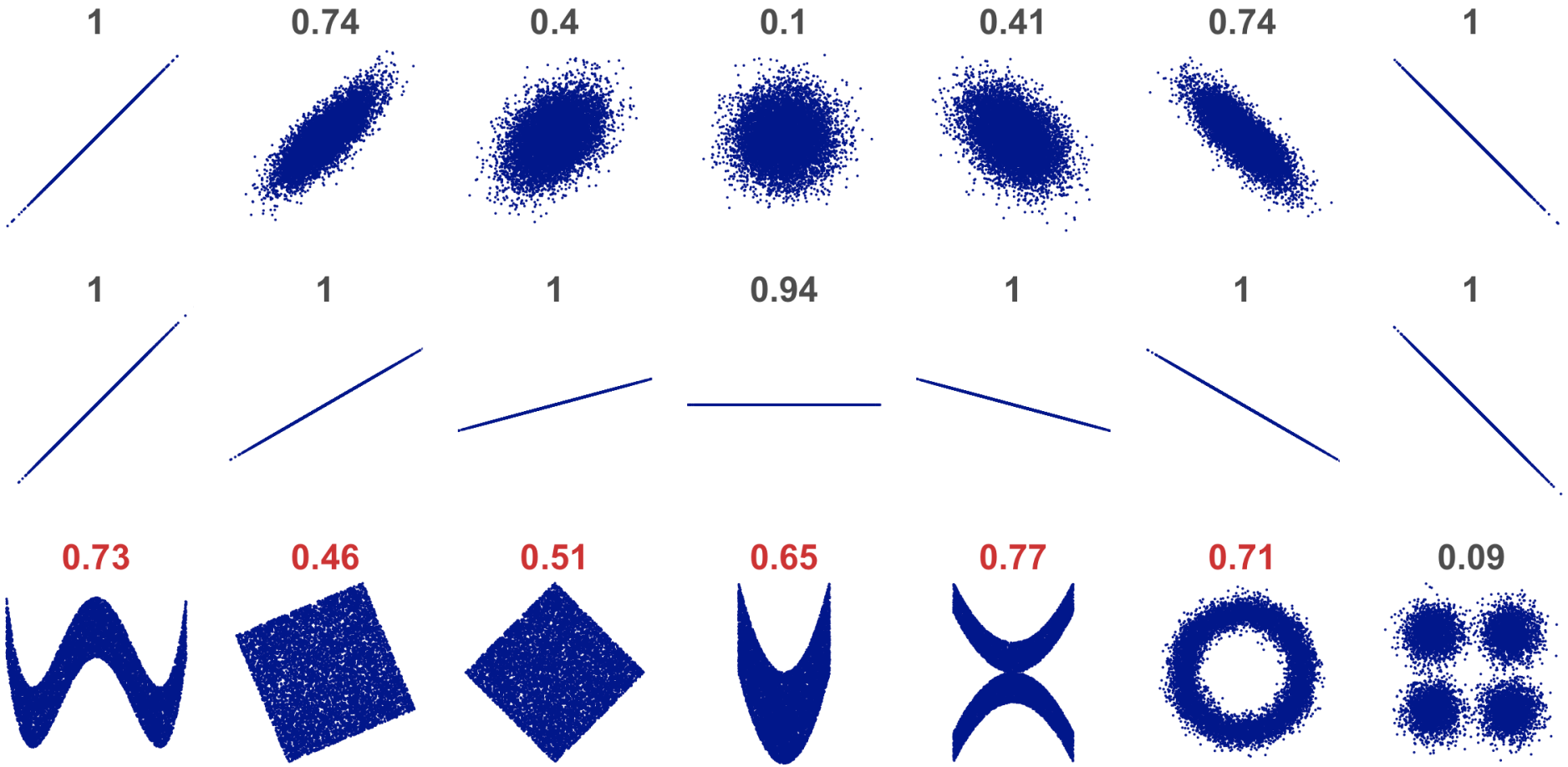
近似バイアスが確認できるものの、非線形パターンの検出は良好



n = 10,000, k = 10　一般化相関尺度

平均サイズ100を基準としてkを設定（$n/k^2 = 100$）

独立に近い例で近似バイアスが低下している



提案手法の評価

- シミュレーションにより理論値に収束する傾向を確認
- 実際のデータにおいても非線形パターンを検出
- ある程度は制御可能だが、近似バイアスが課題
  - 各種のダイバージェンス推定法を応用することで
    改善される可能性も期待される（例：Wang et al., 2005）

一方で以下の点は提案手法のメリットといえる：

- 単調変換に対して不変
- 外れ値の影響を受けにくい
- 離散値や欠損値を含むデータに適用しやすい
- 計算コストが低く、大規模データに適用しやすい

（参考）プラグイン推定量のバイアス

離散化された変数をX_s, Y_tとおき、支持集合の基数をs, tとする

χ^2統計量を$\hat{\chi}^2$と書くと、独立性の帰無仮説のもとで

$$\hat{\chi}^2 \sim \chi^2_{(s-1)(t-1)} \quad (n \to \infty)$$

このとき期待値は$E\left[\hat{\chi}^2\right] \approx (s-1)(t-1)$

プラグイン推定量$\hat{\psi} = \hat{\chi}^2/n + 1$について

$$E\left[\hat{\psi}\right] \approx 1 + \frac{(s-1)(t-1)}{n}$$

帰無仮説のもとで$\psi = 1$

$\Rightarrow$　独立であっても分割数が増えると上方バイアスが生じる

（参考）改善案：保守的なバイアス補正

帰無仮説のもとで（漸近的に）不偏な推定量を定義：

$$\tilde{\psi} := \hat{\psi} - \frac{(s-1)(t-1)}{n}$$

$\psi \geq 1$より、推定値としては$\max\{1, \tilde{\psi}\}$を用いる

> - クラメールのVに関しても類似のバイアス補正が提案されている (Bergsma, 2013)
>   - 正確確率検定（周辺度数を固定）のもとで分母は$n-1$

ソフトウェア実装

Rパッケージ `gcor`

開発版をGitHubにて公開中（オープンソース）

https://github.com/r-suzuki/gcor

- 一般化相関尺度および関連指標を算出
- 数値とカテゴリの両方に対応
- 欠損値や外れ値を含むデータにも有効

サンプルデータ

`iris`データ

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

情報をもつ欠損（informative missingness）

`Species == "setosa"`のとき50%の確率で`Sepal.Width`が欠損

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          NA          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7          NA          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4          NA          1.7         0.4  setosa
```

標本相関係数

カテゴリ値を含むためエラー

```
cor(iris)
```

```
Error in cor(iris): 'x' must be numeric
```

数値変数のみに限定して実行

```
cor(iris[, 1:4])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

一般化相関尺度

数値とカテゴリを統一的に評価

```
gcor(iris)
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
Sepal.Length    1.0000000   0.2349075    0.8846517   0.8741873 0.7623968
Sepal.Width     0.2349075   1.0000000    0.3143301   0.2669031 0.6510740
Petal.Length    0.8846517   0.3143301    1.0000000   0.9503289 0.8221674
Petal.Width     0.8741873   0.2669031    0.9503289   1.0000000 0.8237429
Species         0.7623968   0.6510740    0.8221674   0.8237429 1.0000000
```

分割数kを指定（既定ではサンプルサイズから自動設定）

```
gcor(iris, k = 3)
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
Sepal.Length    1.0000000   0.6180293    0.8449216   0.8300394 0.8226041
Sepal.Width     0.6180293   1.0000000    0.6862189   0.6963419 0.6794982
Petal.Length    0.8449216   0.6862189    1.0000000   0.9581278 0.9728947
Petal.Width     0.8300394   0.6963419    0.9581278   1.0000000 0.9795837
Species         0.8226041   0.6794982    0.9728947   0.9795837 1.0000000
```

標本相関係数（欠損を含むデータ）

欠損を含む列は算出されない

```
cor(iris_NA[, 1:4])
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000          NA    0.8717538   0.8179411
Sepal.Width            NA           1           NA          NA
Petal.Length    0.8717538          NA    1.0000000   0.9628654
Petal.Width     0.8179411          NA    0.9628654   1.0000000
```

列の組ごとに、欠損を含むケースを無視する

```
cor(iris_NA[, 1:4], use = "pairwise.complete.obs")
```

```
             Sepal.Length   Sepal.Width Petal.Length Petal.Width
Sepal.Length 1.0000000000  0.0002584192    0.8717538   0.8179411
Sepal.Width  0.0002584192  1.0000000000   -0.3163403  -0.2364441
Petal.Length 0.8717537759 -0.3163403009    1.0000000   0.9628654
Petal.Width  0.8179411263 -0.2364441194    0.9628654   1.0000000
```

一般化相関尺度（欠損を含むデータ）

「欠損という事象を観測した」ものとして評価

```
gcor(iris_NA)
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
Sepal.Length    1.0000000   0.5017295    0.8846517   0.8741873 0.7623968
Sepal.Width     0.5017295   1.0000000    0.5347735   0.5141064 0.7313361
Petal.Length    0.8846517   0.5347735    1.0000000   0.9503289 0.8221674
Petal.Width     0.8741873   0.5141064    0.9503289   1.0000000 0.8237429
Species         0.7623968   0.7313361    0.8221674   0.8237429 1.0000000
```

欠損を無視することも可能（この例では評価が変わる）

```
gcor(iris_NA, dropNA = "pairwise")
```

```
             Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
Sepal.Length  1.000000000 0.009496501   0.88465174  0.87418733 0.7623968
Sepal.Width   0.009496501 1.000000000   0.09600307  0.04385796 0.5701664
Petal.Length  0.884651737 0.096003072   1.00000000  0.95032889 0.8221674
Petal.Width   0.874187331 0.043857963   0.95032889  1.00000000 0.8237429
Species       0.762396795 0.570166409   0.82216737  0.82374290 1.0000000
```

要求の振り返り：達成できたこと

- 線形のみならず非線形な関係も評価
- データ型や分布を問わず、統一的に適用可能
  - 連続、離散、あるいはその混合
  - 外れ値や欠損を含む
  - 分布が非対称、裾が重い、多峰的
- 測定単位を問わない（単調変換に対して不変）
- 計算コストが低く、大規模データにも適用可能
  - 数値を階級に変換し、組み合わせを集計（分位点でなくてもよい）
  - 実装も容易（加減乗除と平方根のみで推定値を算出可能）

⇒　**Rによるソフトウェア実装を公開し、容易に実行可能**

要求の振り返り：課題が残る部分

データや分析手法の前提知識

- サンプルサイズが小さく非線形な関係があるとき、
  分割数kの設定を検討する必要がある
- 値の評価にあたって近似バイアスを意識する必要がある

統計的推論（仮説検定、信頼区間の構成）

- χ^2分布やブートストラップを用いた推論を期待したが、
  シミュレーション結果から近似バイアスの影響が懸念される
- 現時点では探索的な手法としての位置付けが妥当と考えられる

$\Rightarrow$ **いずれの観点からも近似バイアスが課題**

今後の課題

- 議論の整理
  - 既存手法との比較など、参照すべき先行研究も多数
- 研究成果の公開
  - ソフトウェアとの相乗効果で応用に繋げる
- 手法の改良
  - 近似バイアスを含む推定精度の改善
  - 統計的推論
  - 異なる観点からの一般化
- ソフトウェアの改良
  - 数値実験を想定した冗長な処理を除去
  - 多値カテゴリの取り扱い
  - データ可視化などの応用
  - CRAN登録、Python版の検討など

本日の資料とRパッケージ

https://r-suzuki.github.io/ja