

# 非線形相関の評価指標および推定手法

---

鈴木 了太  
[suzuki@ef-prime.com](mailto:suzuki@ef-prime.com)  
株式会社 ef-prime

2025年11月23日  
統計数理研究所共同利用研究集会 「データ解析環境Rの整備と利用」

# 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

## 量的データ

- 相関係数：線形（または単調）な相関関係を評価
  - 積率相関係数、順位相関係数
- 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）

## 質的データ

- 連関係数：独立性を評価
  - 例：クラメールの $V$ 、ピアソンの $\phi$ 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
  - 連續、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（変換に対して不变）
- 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

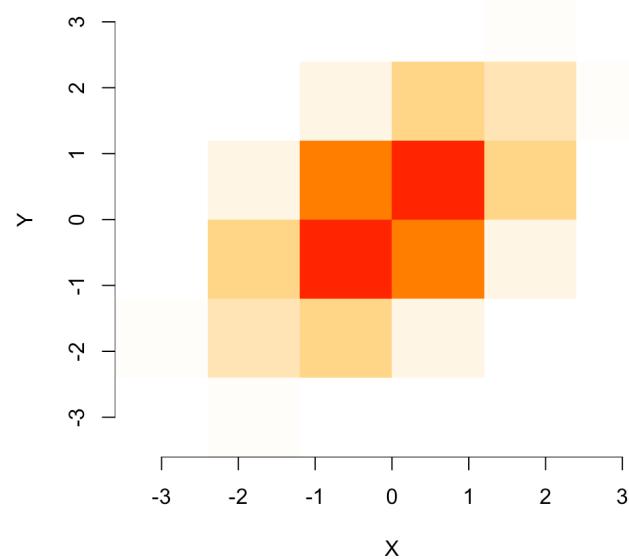
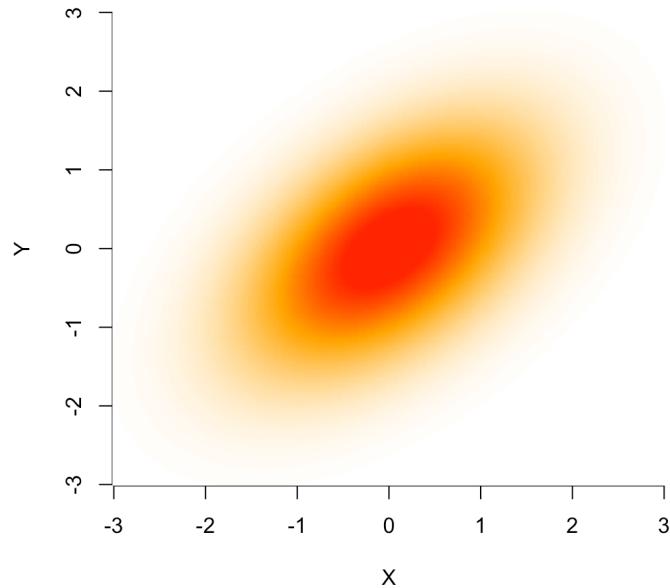
## 指標の設計方針

- $X$ 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：連續変数の離散化

連續変数の値域を分割し、離散変数として評価する



$$f(x, y)$$

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

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

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

# 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

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

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

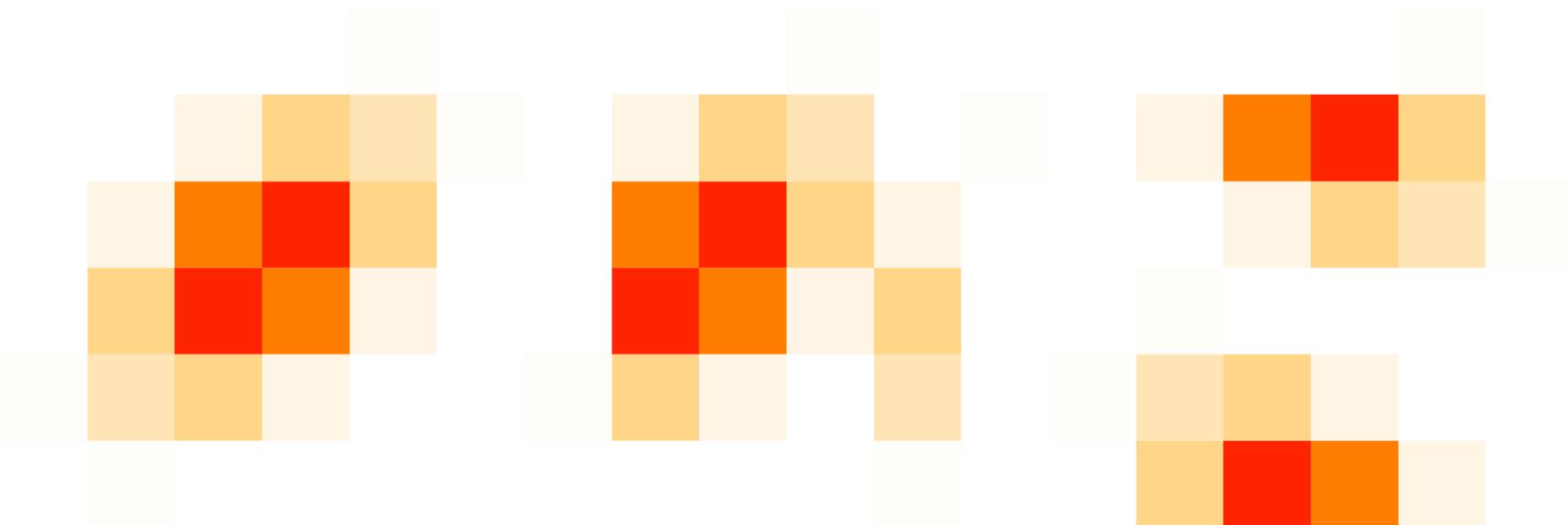
$X$ によって説明できる、 $Y$ の不確実性の割合

- $X$ 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利用 (Suzuki et al., 2006) <https://ef-prime.com/ja/product/>

# 離散化アプローチの利点

- 変数間の関係を2元分割表によって評価
- 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

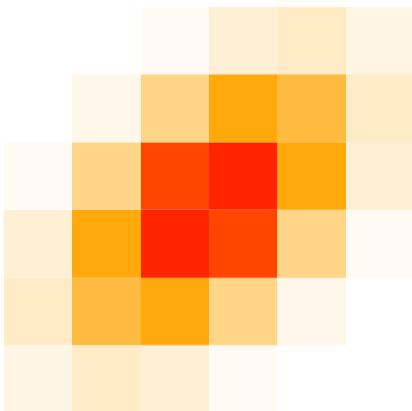


上記の例は相互情報量や不確実性係数の意味で等価

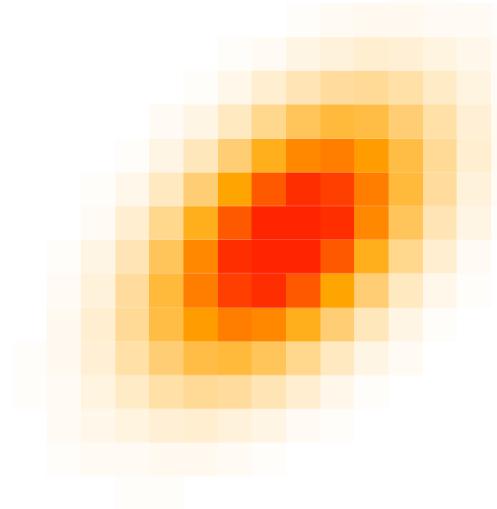
# 連続分布の離散近似

- 変数ごとの分割数  $k$  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

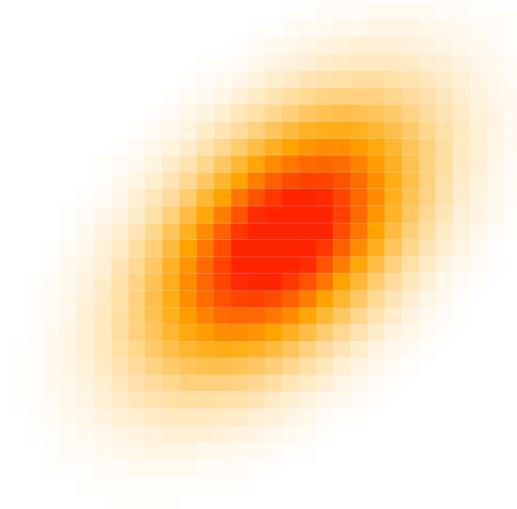
$k = 8$



$k = 16$



$k = 32$



# 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 $\infty$ 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  $I_k \rightarrow I$  ( $k \rightarrow \infty$ )

# 不確実性係数の漸近挙動

不確実性係数  $U(Y_k|X_k)$  について、 $k \rightarrow \infty$  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  $\infty$  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

$$\max_{s,t} \frac{\hat{I}(X_s; Y_t)}{\log(\min\{s, t\})}$$

# 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
  - $k \rightarrow \infty$ 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）

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

$dP/dQ$ 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)

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

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：

$$I(X;Y) = D_{\text{KL}}(P_{X,Y} \| P_X \otimes P_Y)$$

相互依存度はカイ二乗情報量に定数1を足した値。

カイ二乗ダイバージェンスによって定義される：

$$\begin{aligned}\psi(X;Y) &= I_{\chi^2}(X;Y) + 1 \\ &= D_{\chi^2}(P_{X,Y} \| P_X \otimes P_Y) + 1\end{aligned}$$

$\Rightarrow f$ -ダイバージェンスやカイ二乗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（連續、離散あるいは混合）
- 変数変換に対して不变（全单射をなす可測変換）
- 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

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

$(X, Y)$ が相關係数  $\rho$  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、 $\psi$  と  $\rho^2$  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關指標

正規分布における議論に基づいて以下を定義 (\*) :

$$\rho_{\chi^2} = \sqrt{1 - \psi^{-1}}$$

- $(X, Y)$ 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 $\rho_{\chi^2} < 1$

\* 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

$$\rho_I = \sqrt{1 - e^{-2I(X;Y)}}$$

# 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

# 一般化相関尺度

以下を満たす指標を**一般化相関尺度 (Generalized Correlation Measures)**と呼ぶことにし、 $g\text{Cor}(X, Y)$ で表す：

- $[0, 1]$ 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：カイ二乗相関

相互依存度に基づく一般化相関尺度として、

**カイ二乗相関 (Chi-Squared Correlation)** を以下で定義する：

$$r_{\chi^2}(X, Y) := \begin{cases} \sqrt{\frac{1 - \psi^{-1}}{1 - (\sqrt{k_X k_Y})^{-1}}} & (k_X k_Y > 1) \\ 1 & (k_X k_Y = 1) \end{cases}$$

$$\psi = I_{\chi^2}(X; Y) + 1$$

$k_X, k_Y$  は  $X, Y$  の支持集合の基数 (無限集合の場合は  $\infty$ )

- $[0, 1]$  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 变量正規分布に従うとき、相関係数の絶対値  $|\rho|$  と一致

# 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

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

これを  $\psi_k$  と書くと  $\psi_k \leq \psi$

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

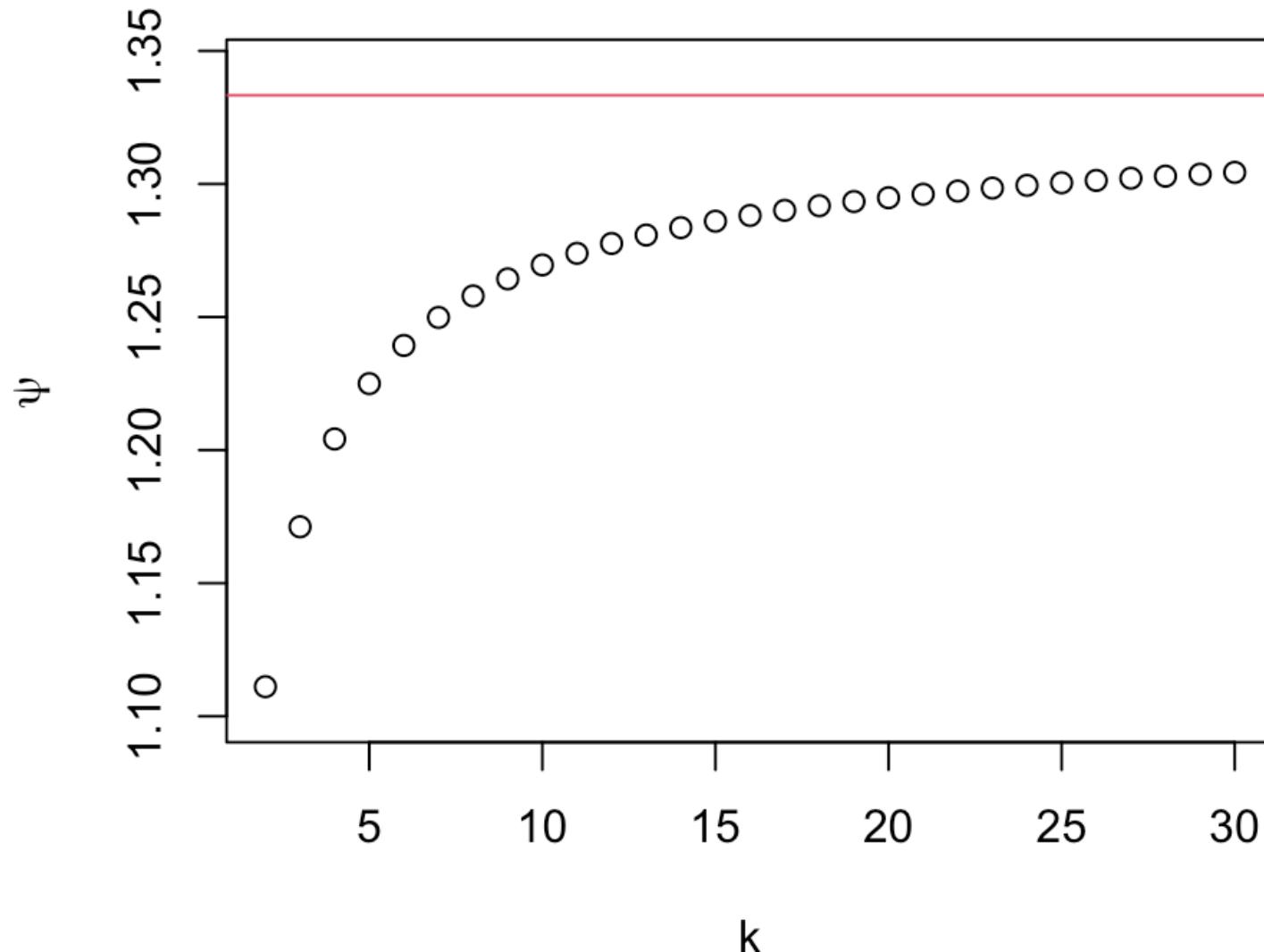
カイ二乗相関についても同様：

$$r_{\chi^2}(X_k, Y_k) \rightarrow r_{\chi^2}(X, Y) \quad (k \rightarrow \infty)$$

# 数値例：相互依存度の近似

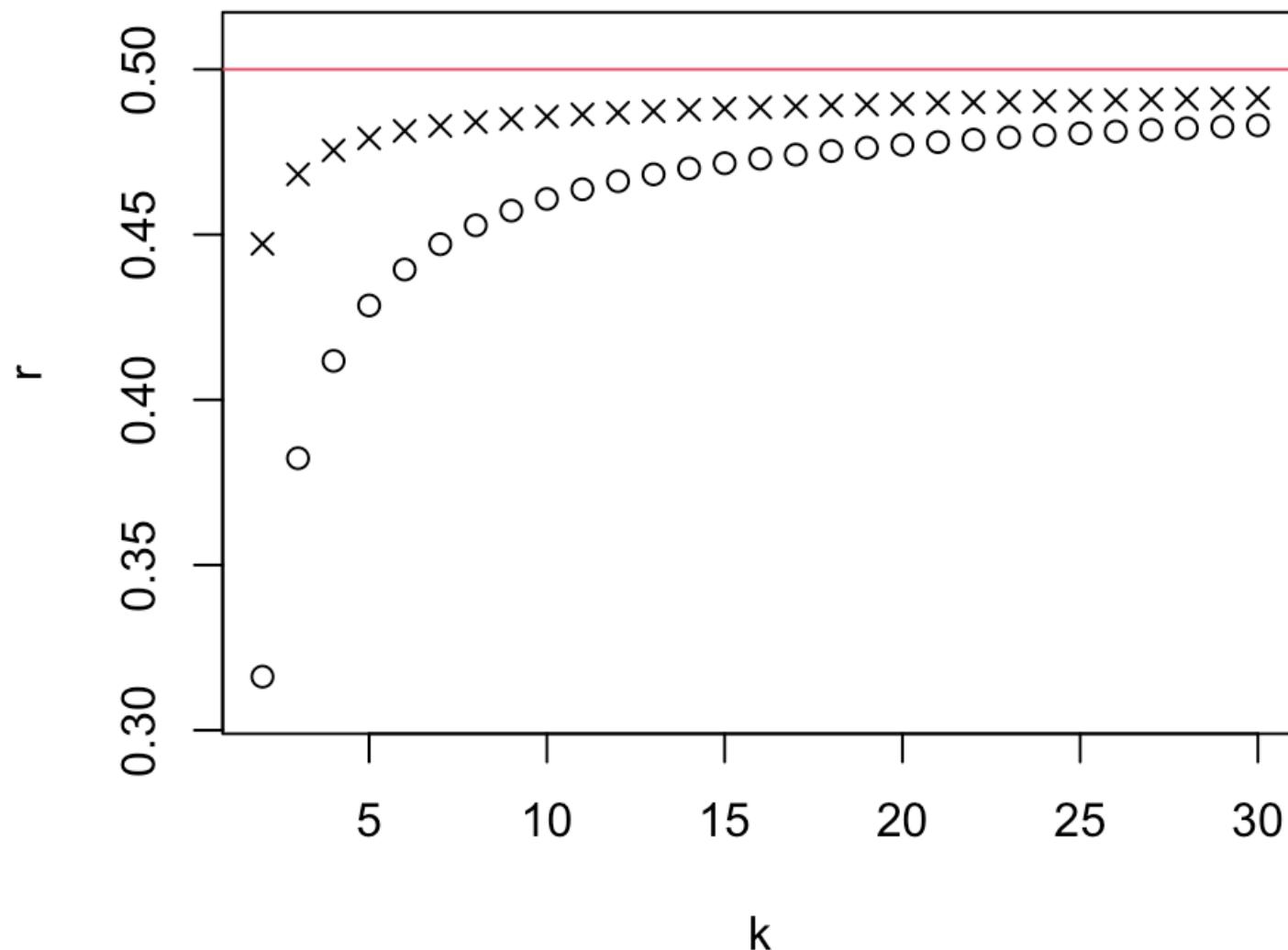
相関係数  $\rho = 0.5$  をもつ2変量正規分布を  $k$  分位点で分割

○が離散化による近似値  $\psi_k$ 、赤線が真値  $\psi = 4/3$



# 数値例：カイ二乗相関の近似

○は離散化による近似値 $\psi_k$ を式 $\sqrt{1 - \psi^{-1}}$ に代入した値  
×が $\sqrt{1 - k^{-1}}$ で基準化した $r_{\chi^2}$ 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、 $r_{\chi^2}(V, W)$ を $r_{\chi^2}(X, Y)$ 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

⇒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： $r_{\chi^2}(V, W) = 0.7$ のとき、背後に $\text{Cor}(X, Y) = 0.7$ 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  $r_{\chi^2}$  などの指標を推定

プラグイン推定量  $\hat{\psi}$  はカイニ乗統計量で表すことができる：

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

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

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

# 連関係数との関係

$X, Y$ が離散のとき、クラメールの $V$ はカイ二乗統計量で表される：

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

$k_X = k_Y = 2$ のときピアソンの $\phi$ 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カイ二乗相関の推定値が得られる：

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

$$\hat{r}_{\chi^2}(X, Y) = \sqrt{\frac{1 - \hat{\psi}^{-1}}{1 - (\sqrt{k_X k_Y})^{-1}}}$$

# 相関係数との関係

$(X, Y)$ が相関係数 $\rho$ 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：

$$\begin{aligned}\sqrt{1 - \hat{\psi}_k^{-1}} &= \sqrt{1 - \left( \frac{\chi^2}{n} + 1 \right)^{-1}} \\ &= \sqrt{\frac{\chi^2/n}{1 + \chi^2/n}} \\ &\rightarrow |\rho| \quad (k, n \rightarrow \infty)\end{aligned}$$

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、 $\hat{r}_{\chi^2}$  はこの値を分割数で基準化したものにあたる

# 推定手法の検討

## 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
2. 分割データについて相互依存度  $\psi$  を推定
3.  $\psi$  の推定値を用いてカイ二乗相関を算出

分割数  $k$  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

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

# 推定シミュレーション

以下を1,000回繰り返す：

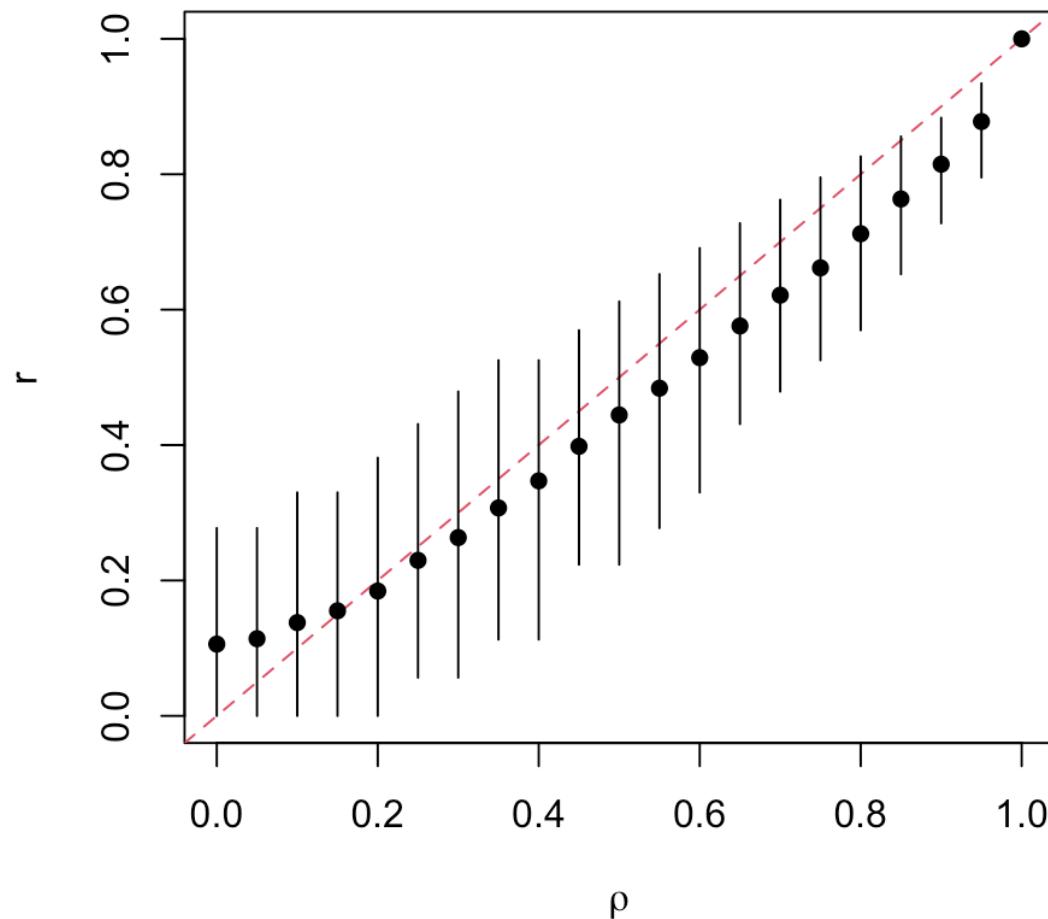
- 平均0、分散1の2変量正規分布に相関係数 $\rho$ を0.05刻みで設定
- サンプルサイズ $n$ のデータを抽出
- 分割数 $k$ を定め、 $\hat{r}_{\chi^2}(X_k, Y_k)$ を計算

⇒ 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

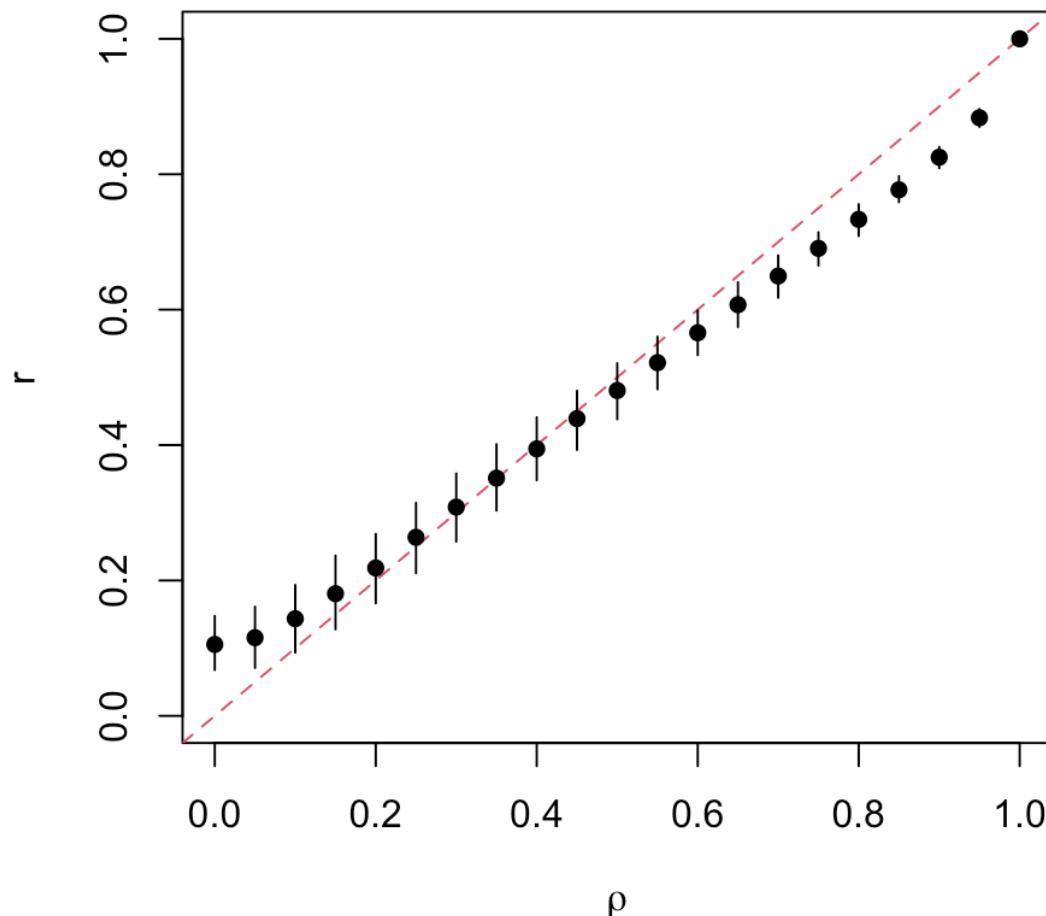
$$n = 100 \quad k = 2$$

$\sqrt{100/50} \approx 1.41 < 2$ を参考に  $k$ を設定、平均サイズ25  
赤の点線は真値。バラツキは大きいが、真値の近くに分布



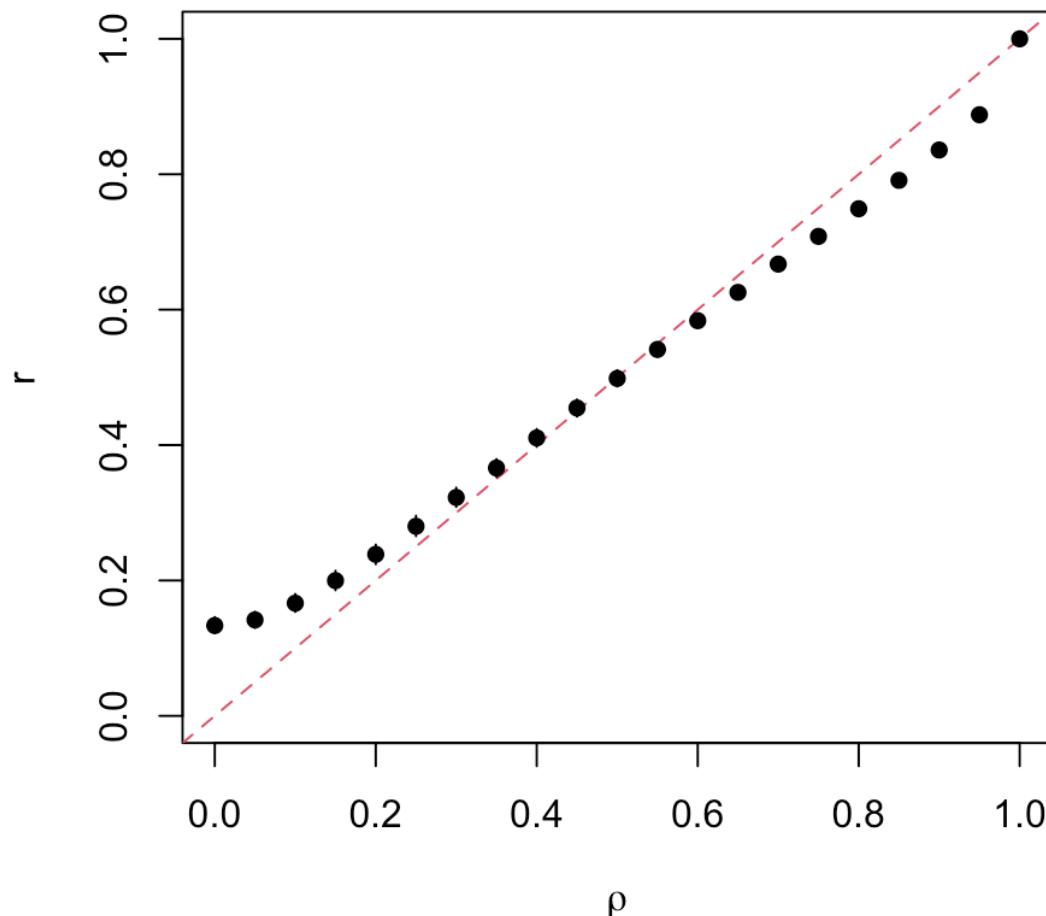
$$n = 1,000 \quad k = 4$$

$\sqrt{1000/50} \approx 4.47$ を参考に  $k$  を設定、平均サイズ 62.5  
バラツキが小さくなるものの、真値からのバイアスは残る



$$n = 10,000 \quad k = 14$$

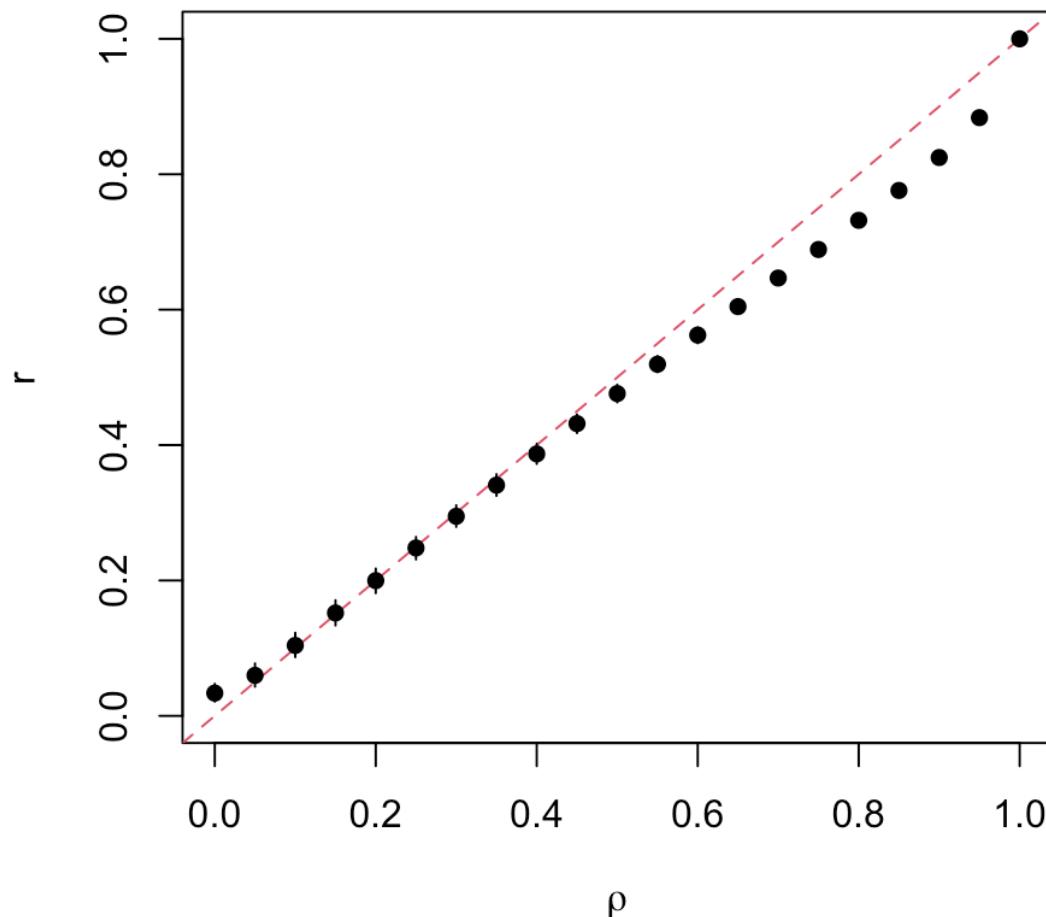
$\sqrt{10000/50} \approx 14.14$ を参考に  $k$ を設定、平均サイズ約51  
バラツキがほとんどなくなるが、バイアスが増加するケースも



$$n = 10,000 \quad k = 4$$

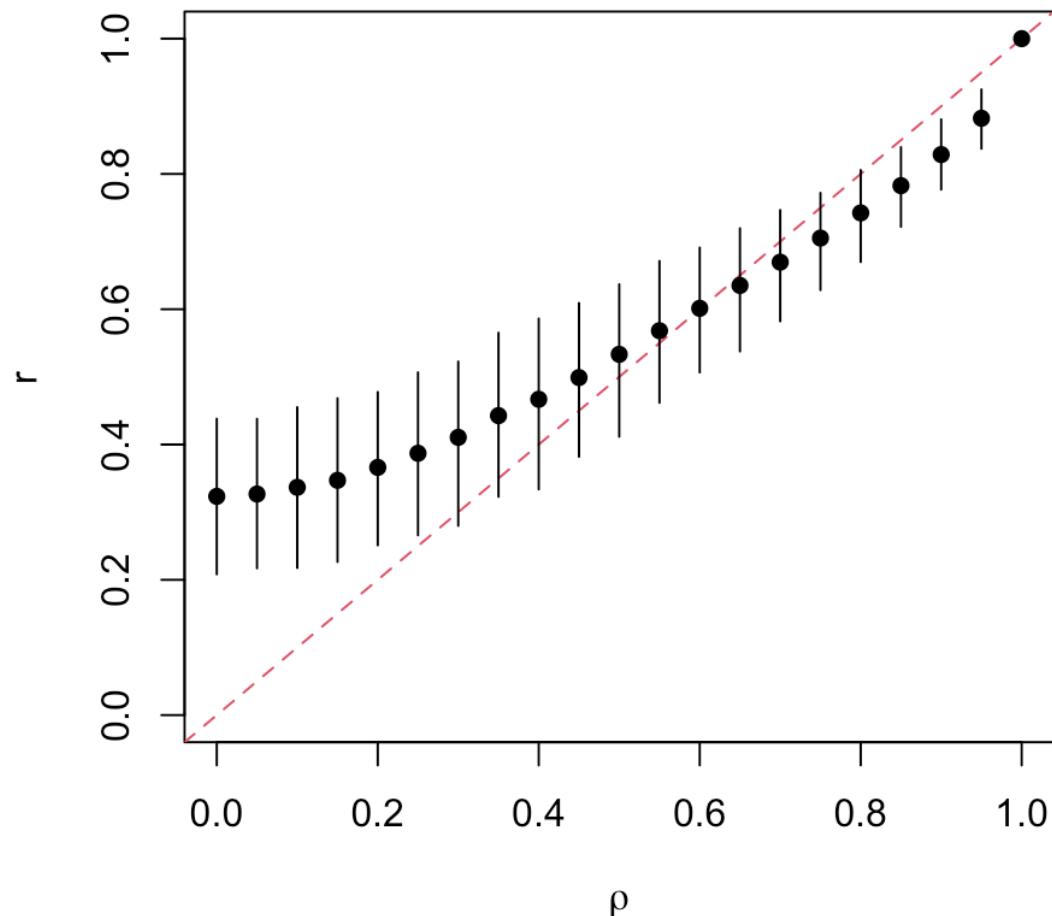
サンプルサイズを固定し、分割数を減らした例

平均サイズが625に増加し、小さい $\rho$ ではバイアスが改善



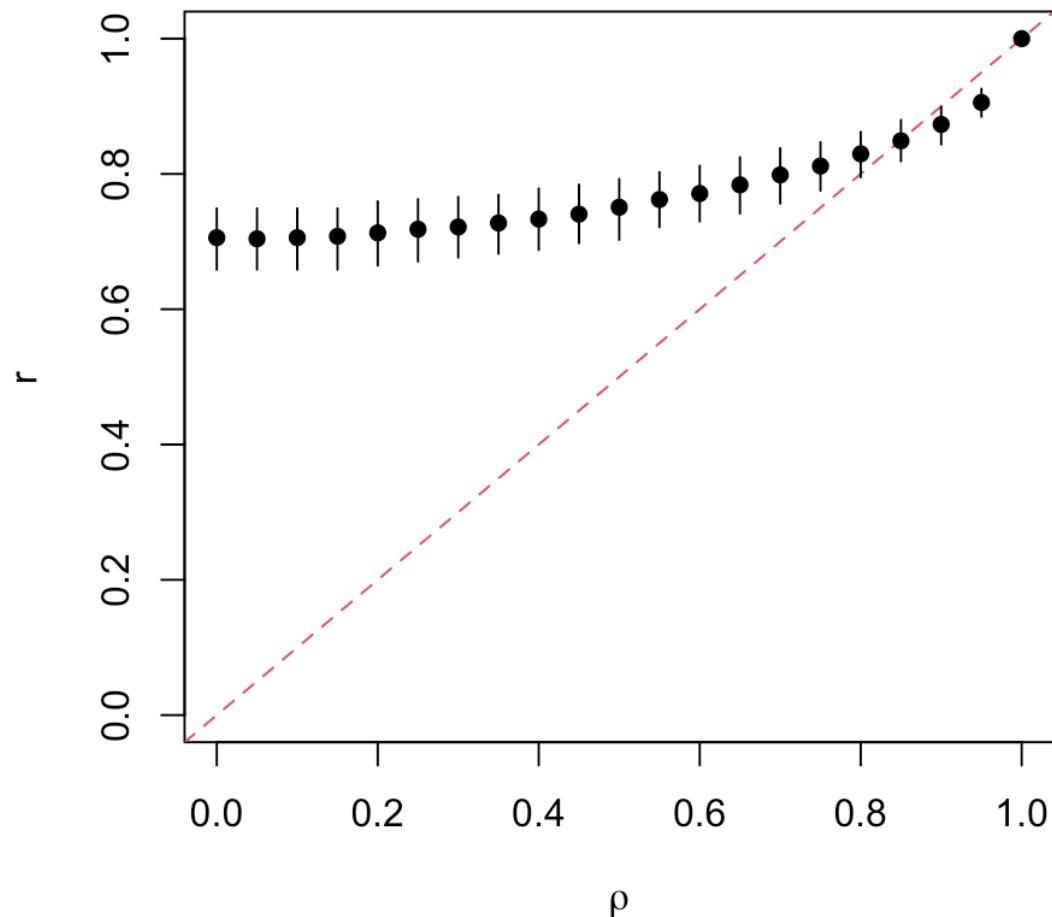
$$n = 100 \quad k = 4$$

分割数を固定し、サンプルサイズを小さくした例。平均サイズは6.25  
バイアスは大きくなるが、順序関係は平均的に保たれている



$$n = 100 \quad k = 10$$

分割数を増やし、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、バラツキは小さくなる



# 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
- 平均サイズ50では $n$ を増やしてもバイアスが残る
- バイアスが大きい場合でも、順序関係は保たれる傾向

⇒ 分割数 $k$ の選択が課題

# 検討：分割数の選択

最適な分割を探索するアプローチも考えられるが、計算コストが高い

以下の観点から検討：

- 推定量の一致性（理論モデルによる考察）
- 有限標本での推定精度（数値シミュレーション）

⇒ 大標本で真値に収束し、小標本でも良好な選択基準を定める

# 分割数の自動選択

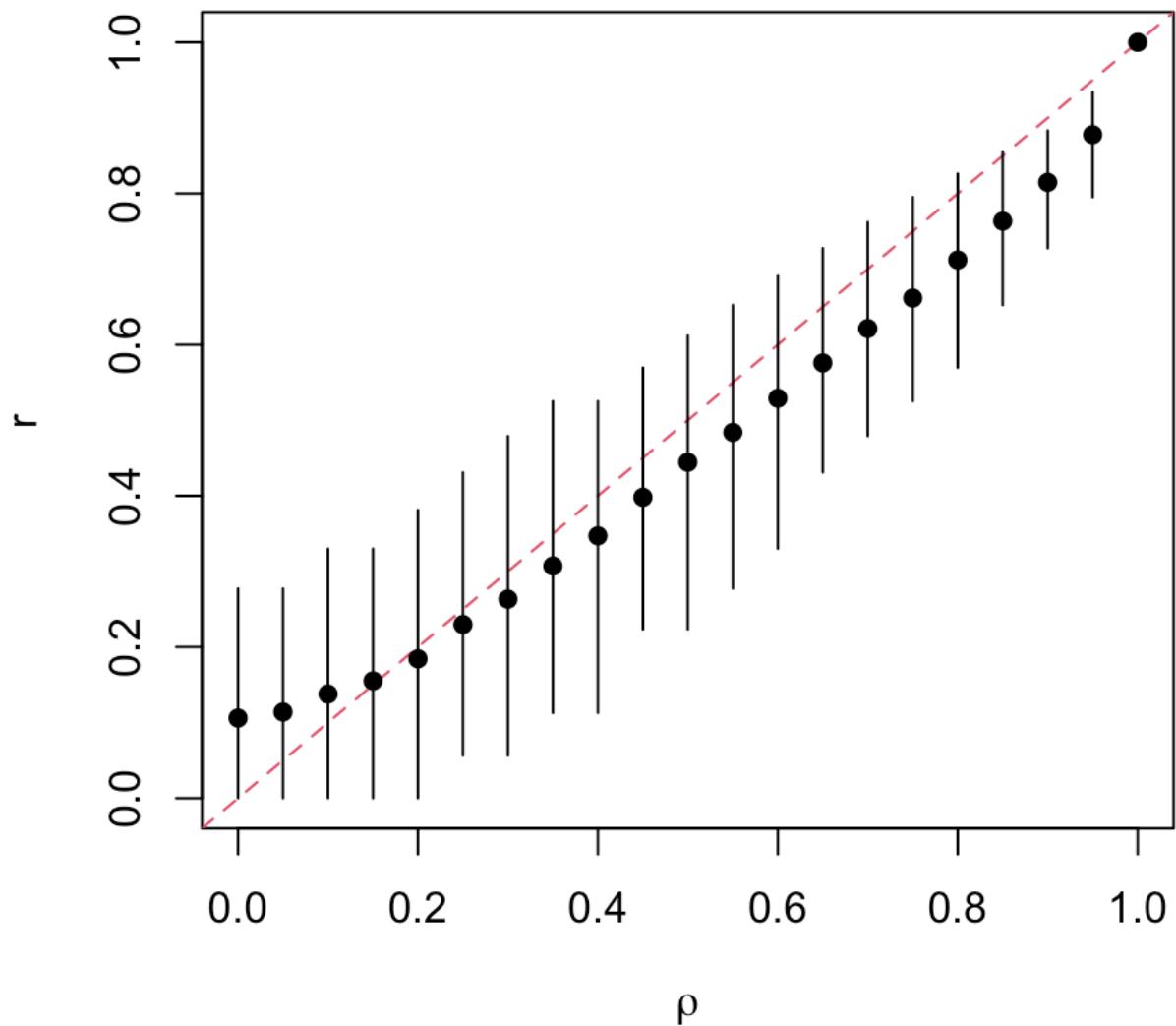
## 現行案：10対2ルール

$$k = \max \left\{ 2, \left\lfloor \frac{1}{2} n^{\log_{10} 2} \right\rfloor \right\}$$

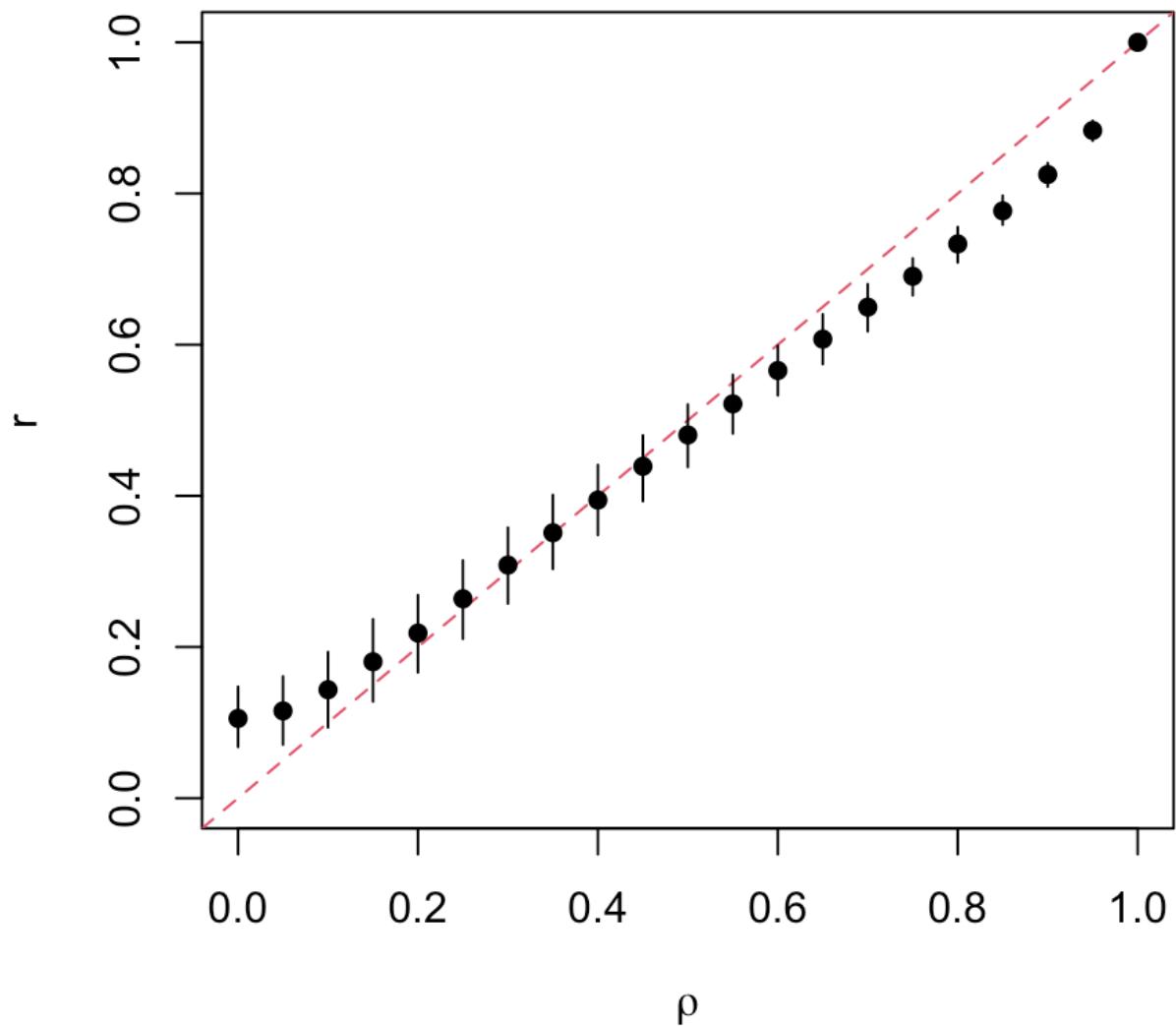
- $n \leq 100$  のとき2分割、 $n$ が10倍のとき $k$ は2倍
- 数値シミュレーションで良好な結果
- 近似モデルに基づく理論上の検討とも整合
  - $k = o(n^{1/2})$  のとき  $n \rightarrow \infty$  のもとで一致性が得られる

$n$	$k$	$n$	$k$
100	2	1,000,000	32
1,000	4	10,000,000	64
10,000	8	100,000,000	128
100,000	16	1,000,000,000	256

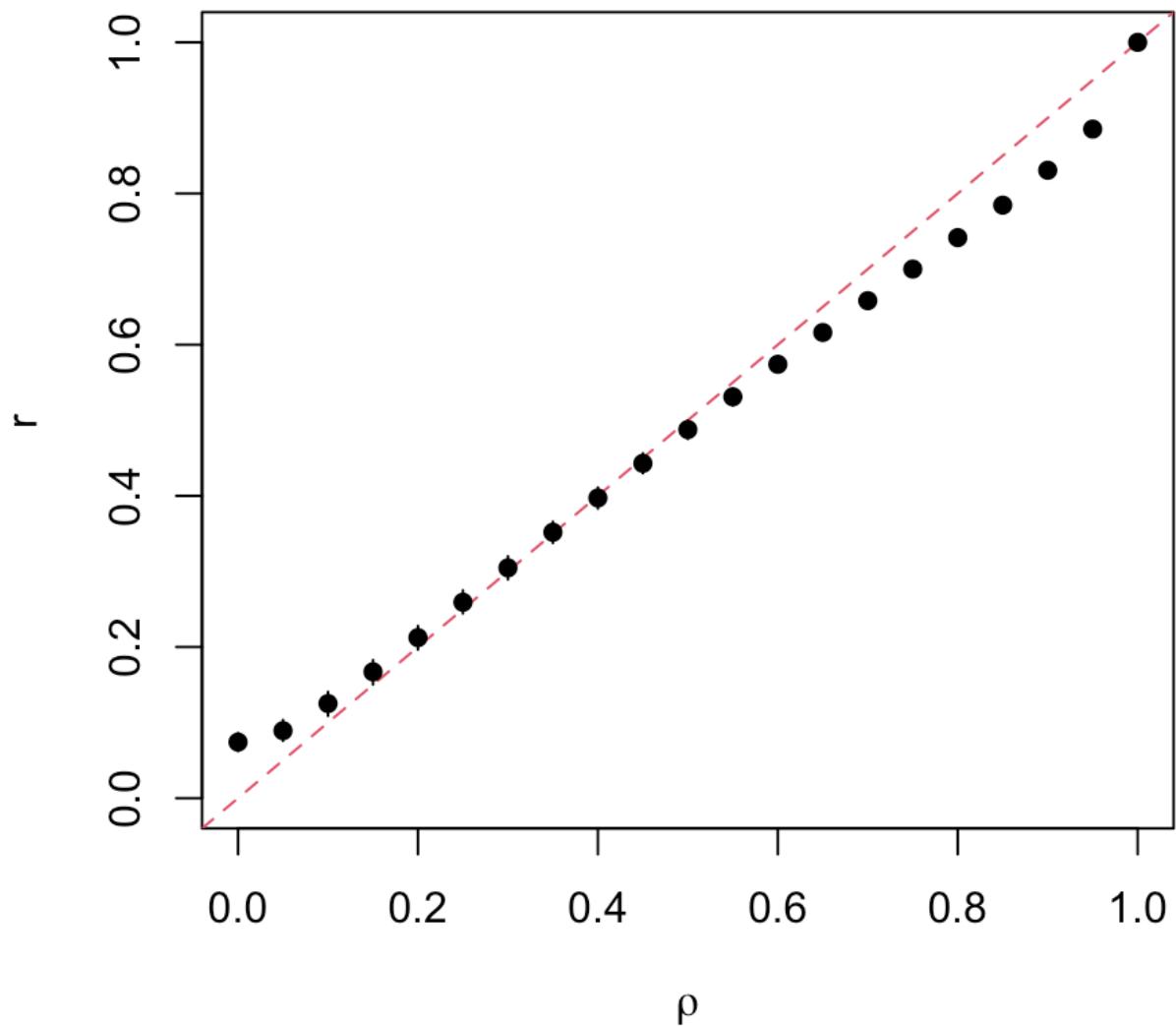
|  $n = 100$     $k = 2$



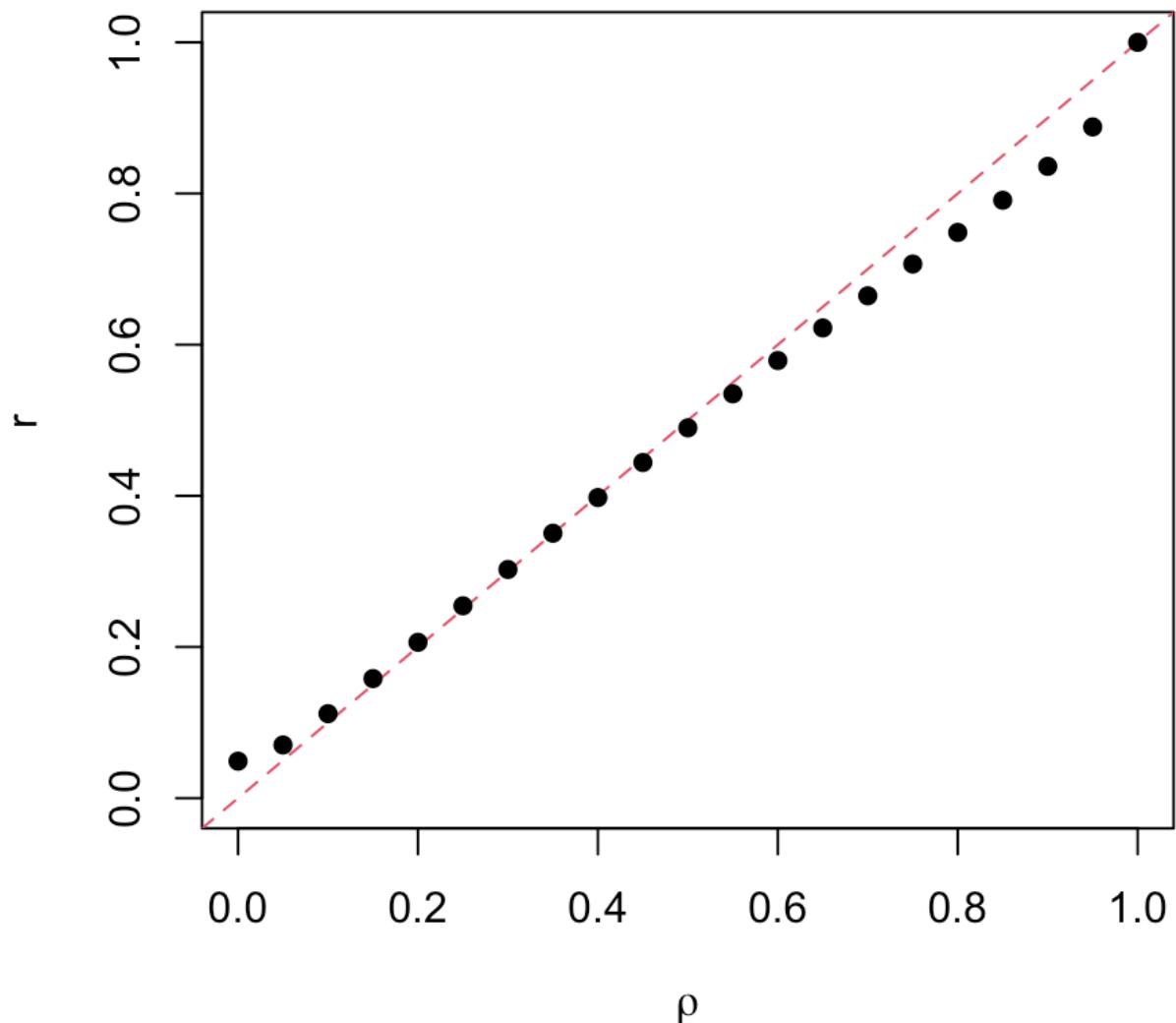
$n = 1,000$      $k = 4$



$n = 10,000$     $k = 8$



$n = 100,000$      $k = 16$



# 提案指標

ここまで検討をもとに、新しい相関指標としてまとめる：

## 分割相関 (contingency correlation)

1. 数値データを標本分位点で分割し、すべての変数 $X_i$ を離散化
  2. 離散化データを用いて相互依存度 $\psi(X_i; X_j)$ を推定
  3.  $\psi$ の推定値を用いてカイ二乗相関 $r_{\chi^2}$ を推定
- 分割数 $k$ は「10対2ルール」などで自動選択、または任意に指定
  - サンプルサイズ $n$ の増加に対して、計算量の増加が緩やか
- ⇒ 大規模データに適したスケーラブルな指標

名称はK. Pearson (1904)のfirst coefficient of contingencyに由来

# 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

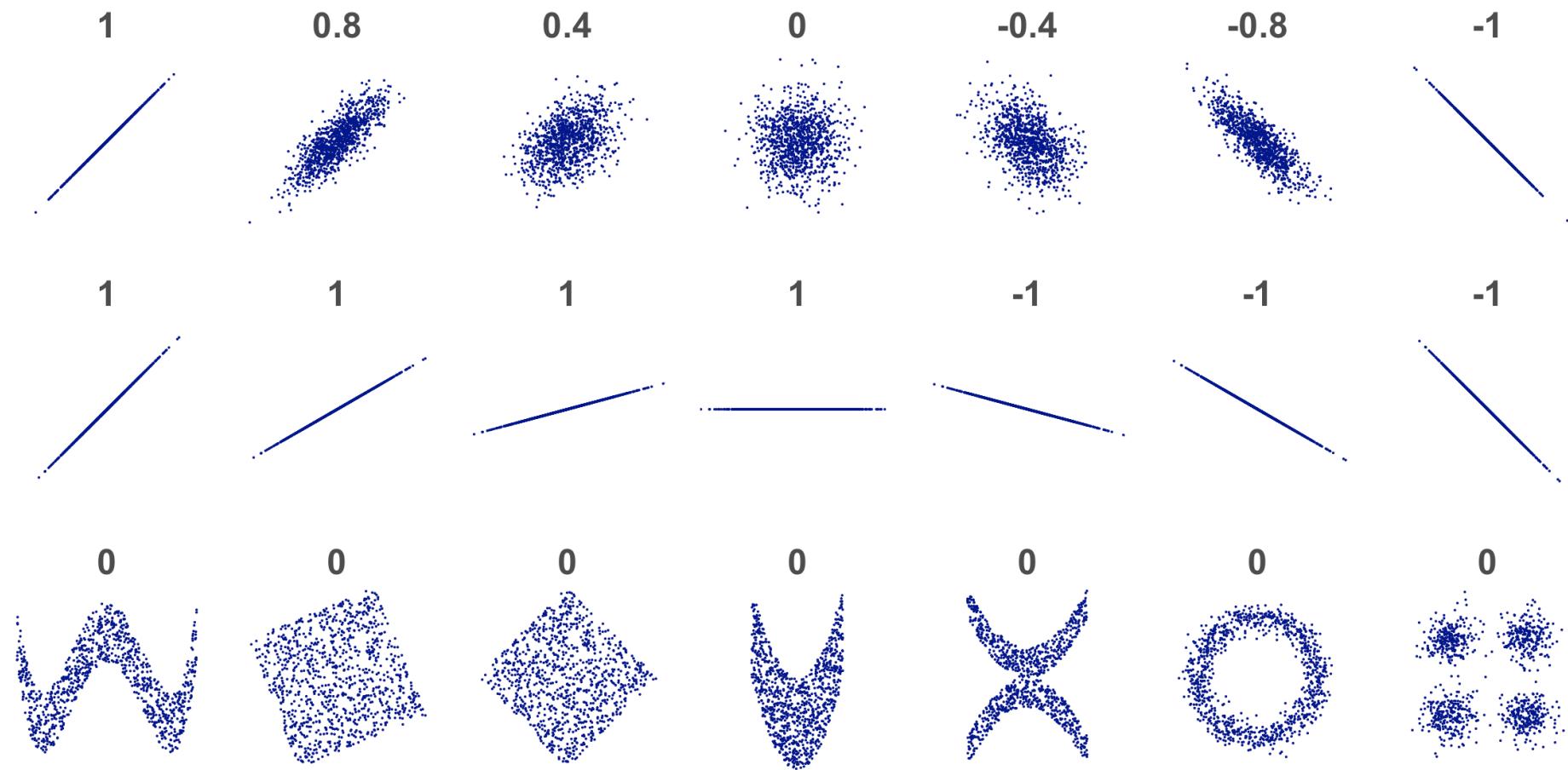
標本相関係数と比較

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

# $n = 1,000$ 標本相關係数

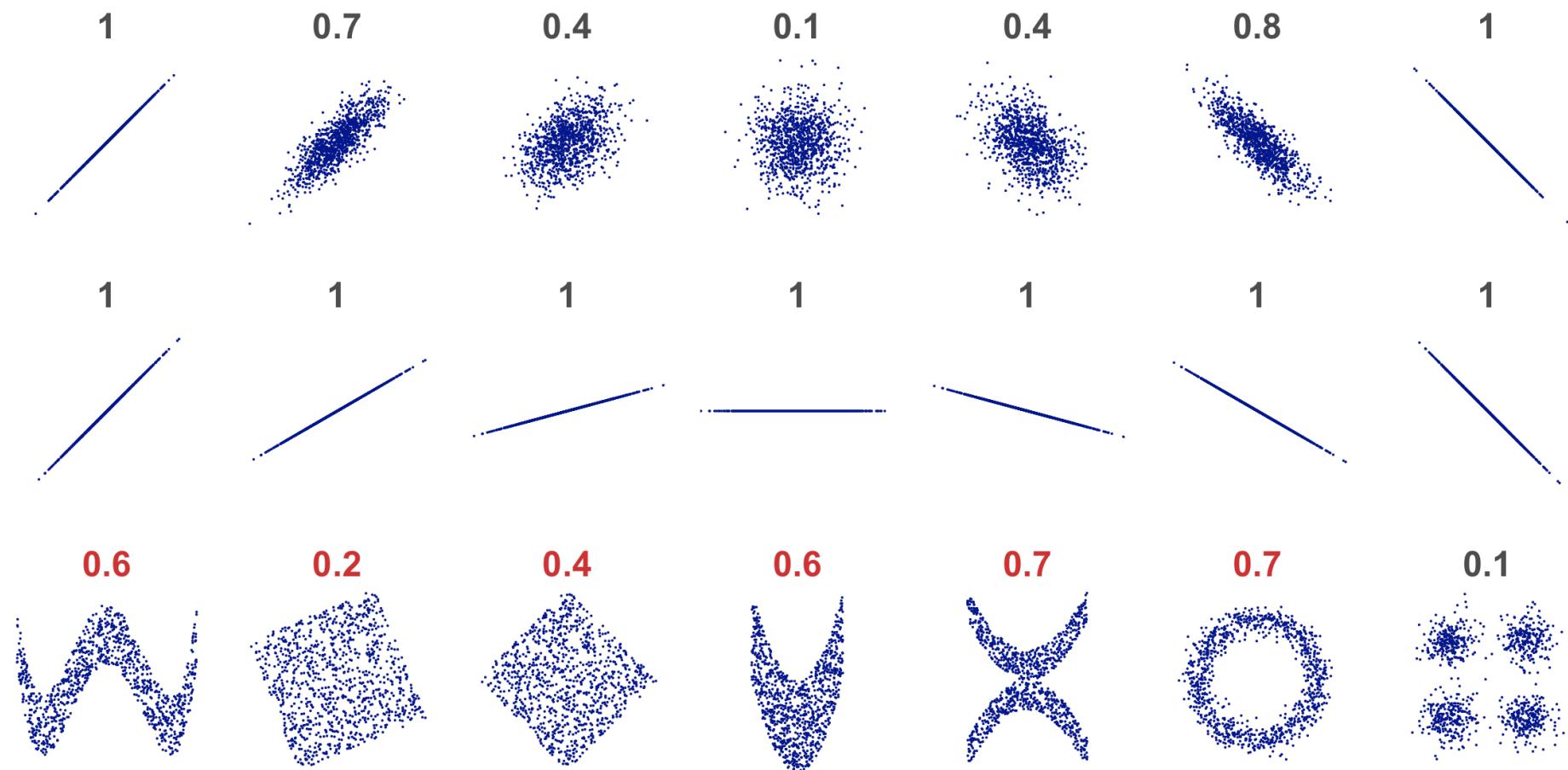
上段は相関係数 $\rho$ の絶対値を $1, 0.8, 0.4, 0$ に設定した2変量正規分布  
小数点以下は1桁に丸めて表示（以降も同様）



# $n = 1,000$ 分割相関 ( $k = 4$ )

10対2ルールに基づいて  $k$  を設定 ( $n/k^2 = 62.5$ )

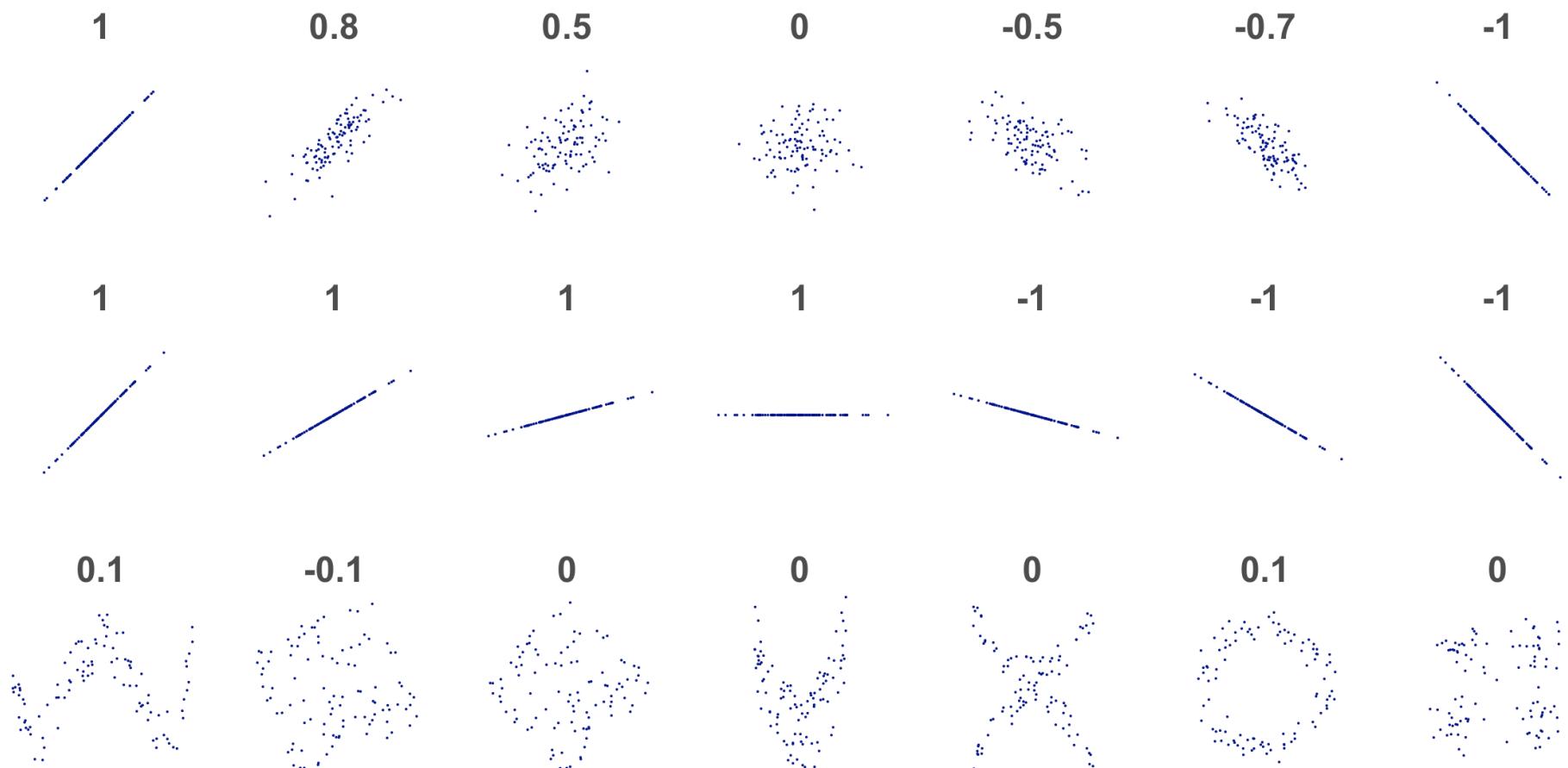
非線形なパターンを検出しつつ、正規分布に対しては  $|\rho|$  に近い値



# $n = 100$ 標本相關係数

サンプルサイズが小さい場合

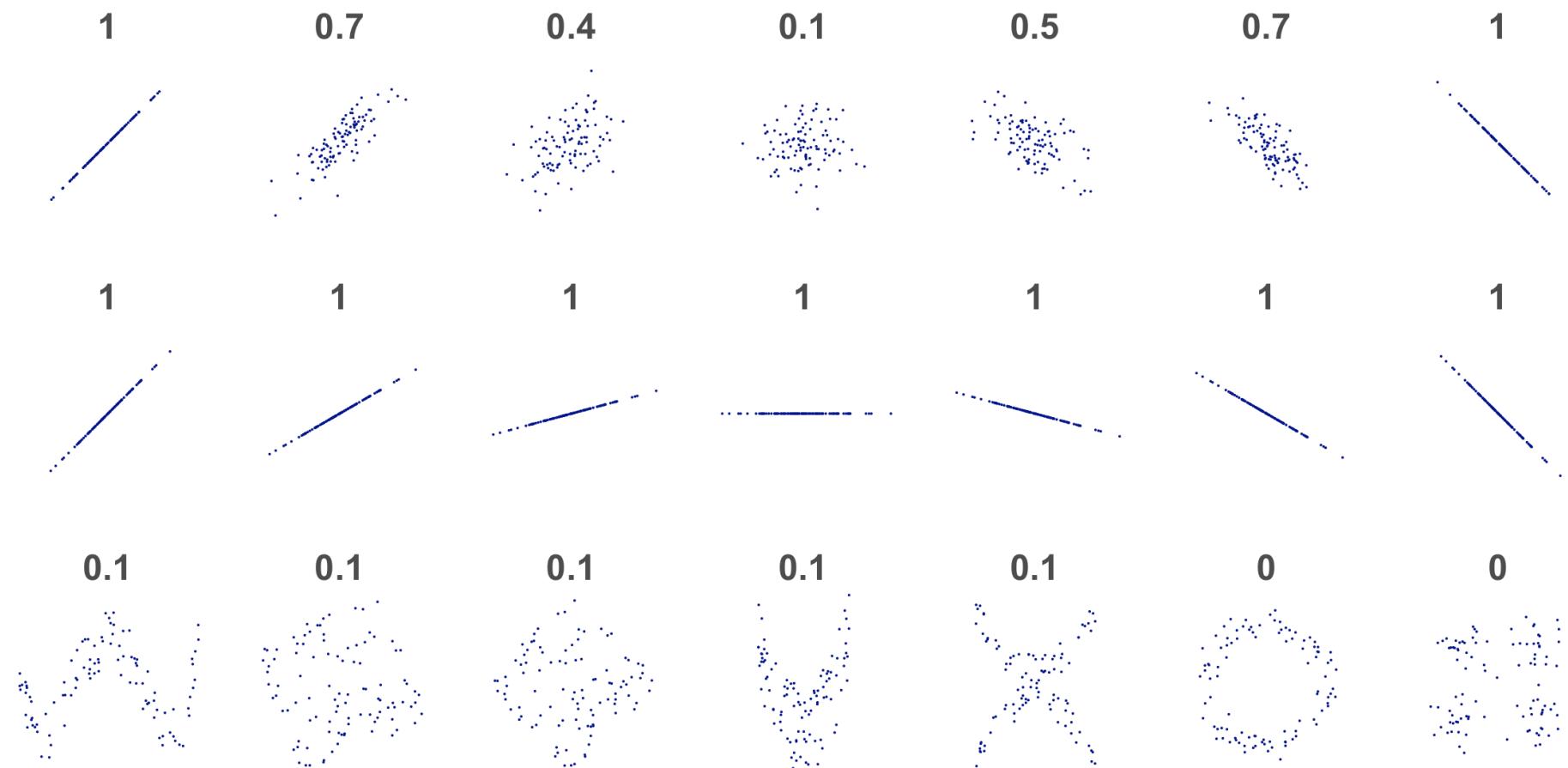
正規乱数に対しても多少の推定誤差が生じる



# $n = 100$ 分割相関 ( $k = 2$ )

10対2ルールに基づいて  $k$  を設定 ( $n/k^2 = 25$ )

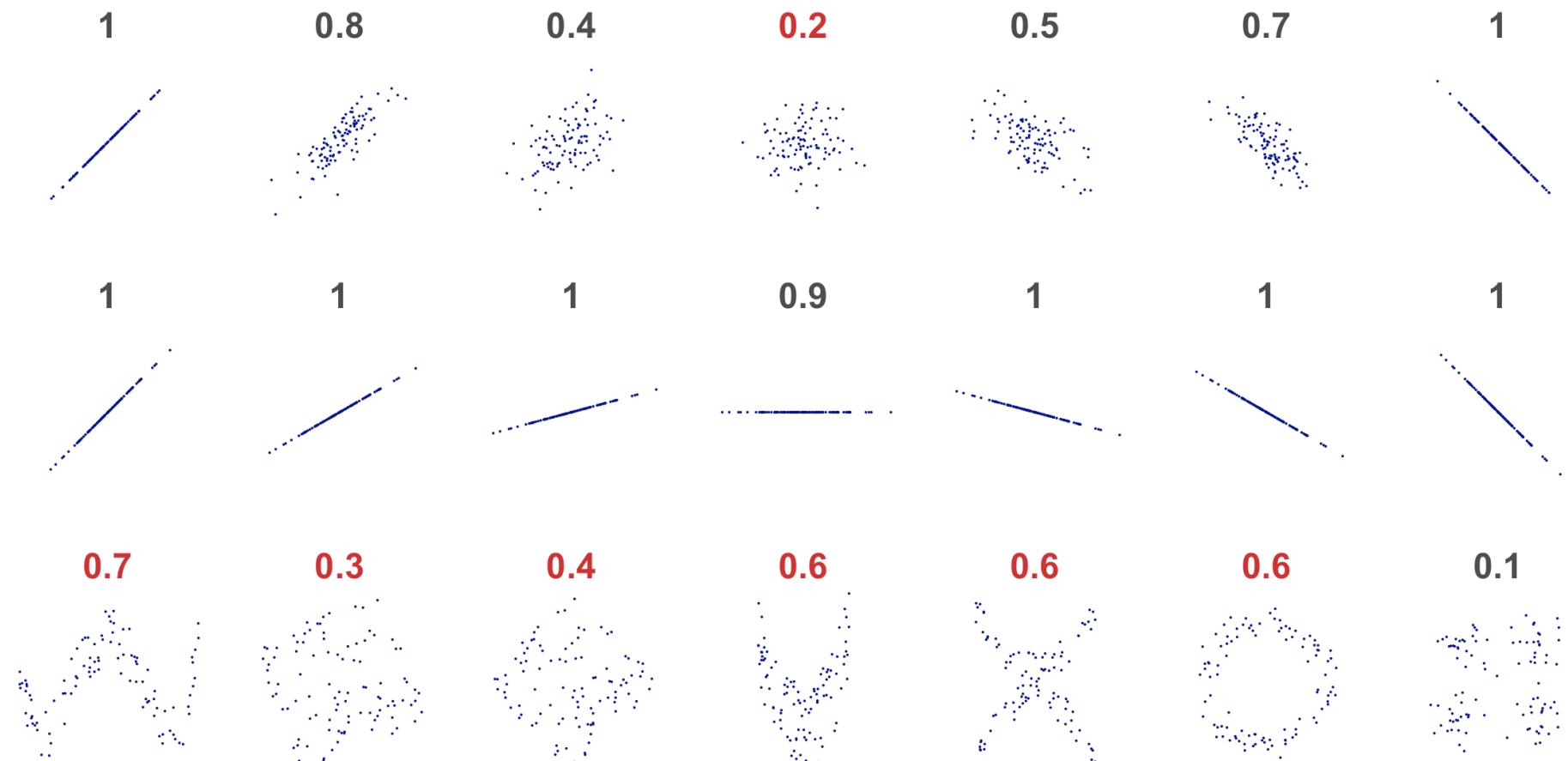
正規分布に対しては  $|\rho|$  に近いが、非線形パターンを検出できていない



# $n = 100$ 分割相関 ( $k = 3$ )

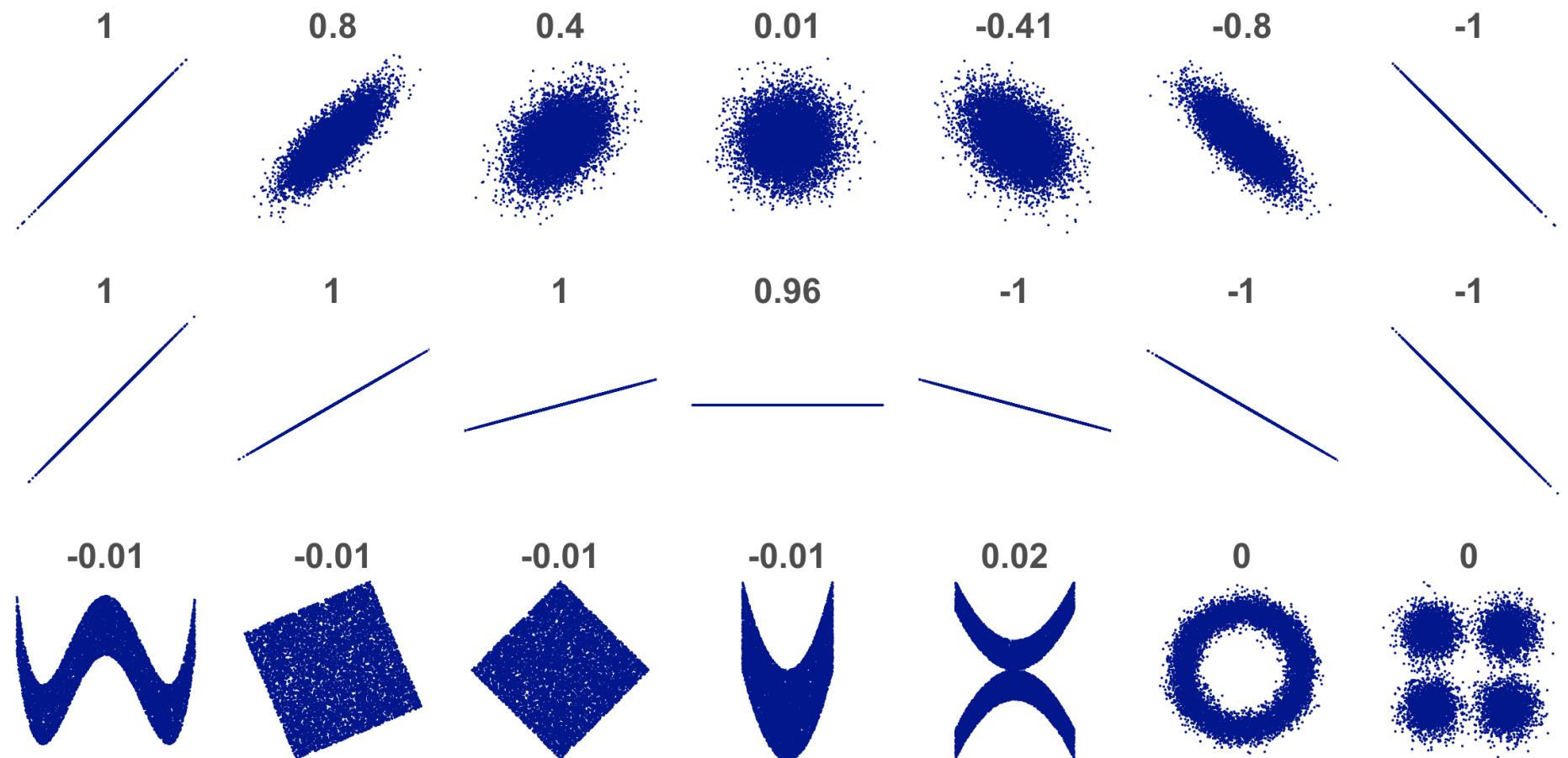
平均サイズ50を大幅に下回る ( $n/k^2 \approx 11$ )

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



# $n = 10,000$ 標本相關係数

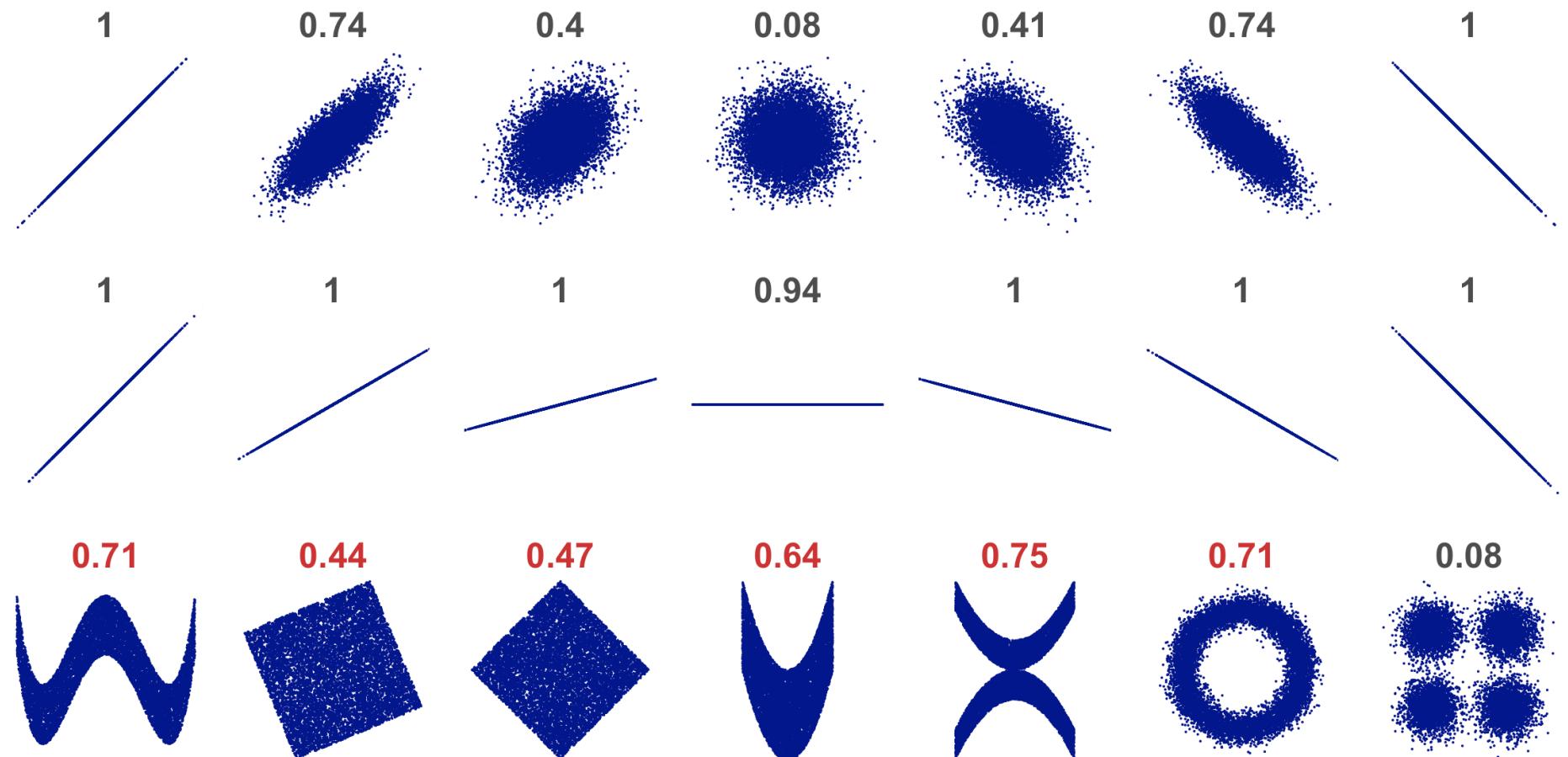
サンプルサイズが大きい場合について、小数点以下第2位まで表示  
サンプルサイズを増やしても、非線形パターンについては0に近い



# $n = 10,000$ 分割相関 ( $k = 8$ )

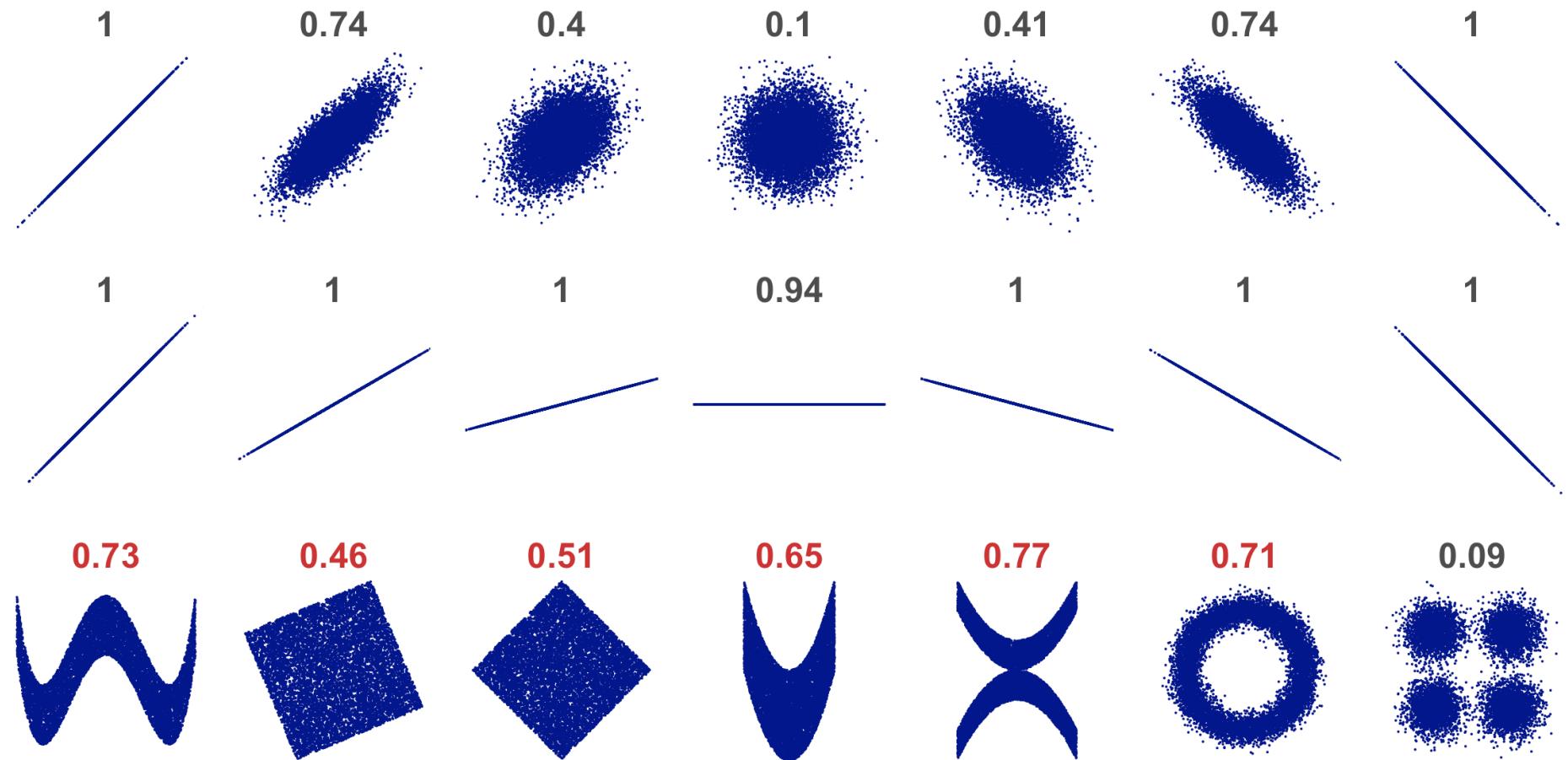
10対2ルールに基づいて  $k$  を設定 ( $n/k^2 \approx 156$ )

非線形パターンの検出は良好、近似バイアスも抑えられている



# $n = 10,000$ 分割相関 ( $k = 10$ )

10対2ルールを参考に、集計結果が解釈しやすい $k$ を設定 ( $n/k^2 = 100$ )  
近似バイアスが若干増加するが、全体の傾向は大きく変わらない



# 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
- ・近似バイアスが課題
  - 「10対2ルール」などで一定の制御が可能
  - 特に小標本で非線形関係の検出性能とトレードオフ

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变
- ・外れ値の影響を受けにくい
- ・数値やカテゴリの混合データに適用可能
- ・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推定バイアス

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  
カイ二乗統計量を  $\hat{\chi}^2$  と書くと、独立性の帰無仮説のもとで

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

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

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

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

帰無仮説のもとで真値は  $\psi = 1$

⇒ 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 $n - 1$

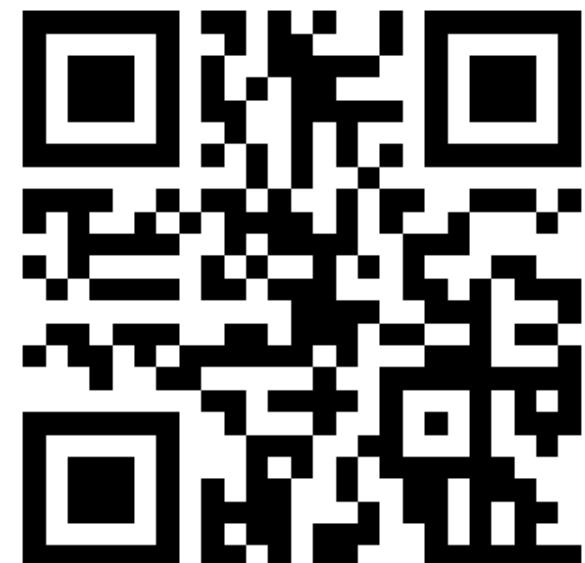
# ソフトウェア実装

## 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>

- 分割相関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	1.000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.000000

# 分割相関

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	1.000000	0.2349075	0.8846517	0.8741873	0.7623968
Sepal.Width	0.2349075	1.000000	0.3143301	0.2669031	0.6510740
Petal.Length	0.8846517	0.3143301	1.000000	0.9503289	0.8221674
Petal.Width	0.8741873	0.2669031	0.9503289	1.000000	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	1.000000	0.6180293	0.8449216	0.8300394	0.8226041
Sepal.Width	0.6180293	1.000000	0.6862189	0.6963419	0.6794982
Petal.Length	0.8449216	0.6862189	1.000000	0.9581278	0.9728947
Petal.Width	0.8300394	0.6963419	0.9581278	1.000000	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		1	NA	NA
Petal.Length	0.8717538		1.0000000	0.9628654
Petal.Width	0.8179411		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	1.000000000	0.0002584192	0.8717538	0.8179411
Sepal.Width	0.0002584192	1.000000000	-0.3163403	-0.2364441
Petal.Length	0.8717537759	-0.3163403009	1.0000000	0.9628654
Petal.Width	0.8179411263	-0.2364441194	0.9628654	1.0000000

# 分割相関（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	1.00000000	0.009496501	0.88465174	0.87418733	0.7623968
Sepal.Width	0.009496501	1.00000000	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
  - 連續、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变）
- 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
  - 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、容易に利用可能に

# 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）

- カイ二乗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

# 今後の課題

- 議論の整理
  - 既存手法との比較など、参考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
  - 推定精度の改善（近似バイアス、推定バイアスなど）
  - 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>



# 補足

---

# 有向一般化相関

以下を満たす指標を**有向一般化相関 (Directed Generalized Correlation)**と呼ぶことにし、 $gCor(X \rightarrow Y)$ で表す：

- $[0, 1]$ 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

特に区別が必要な場合、通常の一般化相関 $gCor(X, Y)$ を  
**無向一般化相関 (Undirected Generalized Correlation)** と呼ぶ

# 提案：有向カイ二乗相関

相互依存度に基づく有向一般化相関として、

**有向カイ二乗相関 (Directed Chi-Squared Correlation)** を以下で定義する：

$$r_{\chi^2}(X \rightarrow Y) := \begin{cases} \sqrt{\frac{1 - \psi^{-1}}{1 - k_Y^{-1}}} & (k_Y > 1) \\ 1 & (k_Y = 1) \end{cases}$$

$$\psi = I_{\chi^2}(X; Y) + 1$$

$k_Y$  は  $Y$  の支持集合の基數 (無限集合の場合は  $\infty$ )

- $[0, 1]$  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