数学と
データサイエンス

～相互活用について

お茶の水女子大学附属
高等学校　数学科
三橋一行

データサイエンスに数学の活用を

相関係数

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \quad \cdots\cdots \quad (1)$$

授業の中で

疑問①　なぜ相関係数ρは －１以上で１以下になるのか。
＜生徒＞

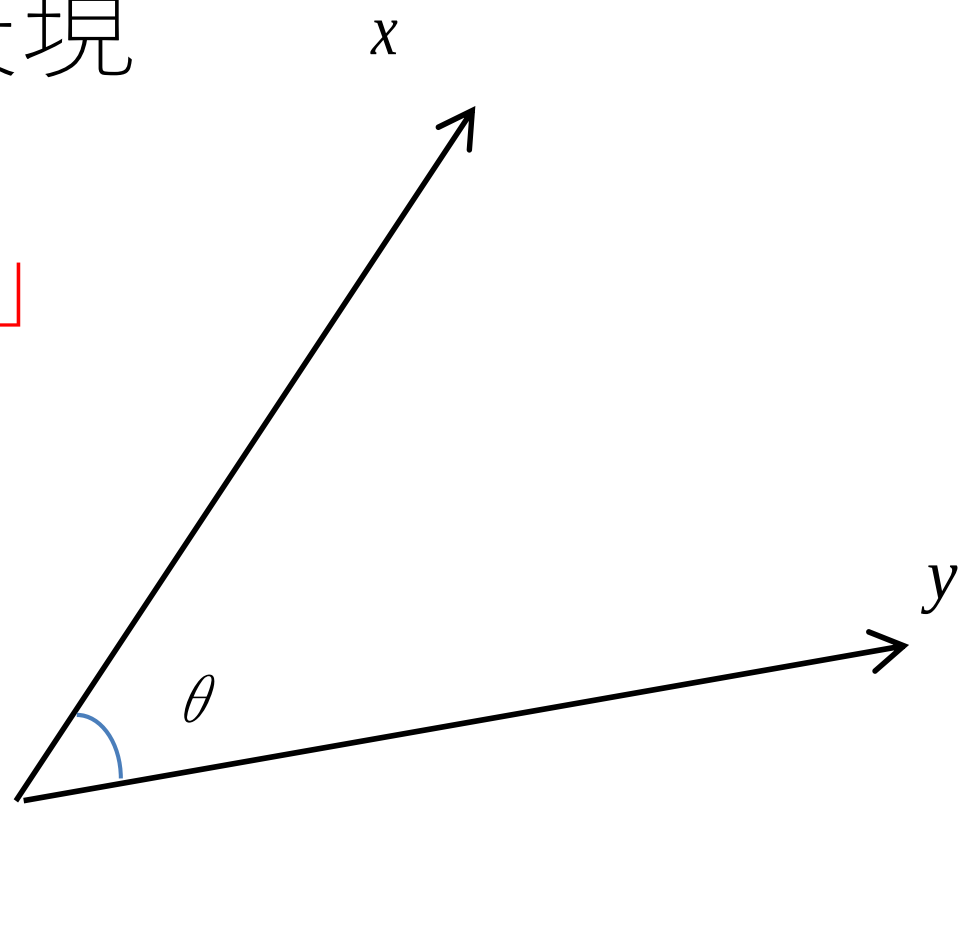
疑問②　相関係数が0.2から0.４になると2倍相関関係が強いといえるのか。＜生徒・教員＞

疑問③　集めたデータから求めた相関係数係を一般化してよいのか。＜教員＞
例：ある国で集められたデータから乳幼児死亡率と母親の識字率の相関関係が高いことがわかったら、それは世界的に（一般的に）認められることなのか。

疑問① ②の解決策として
データをベクトルで表現

- 相関係数とは、
  どれだけ「同じ方向を向いているか」
  を表す尺度。

- サンプルサイズが次元となり、
  次元が高くなるが、ベクトル
  は、矢印で考えると次元の
  ハードルはそれほど高くない。



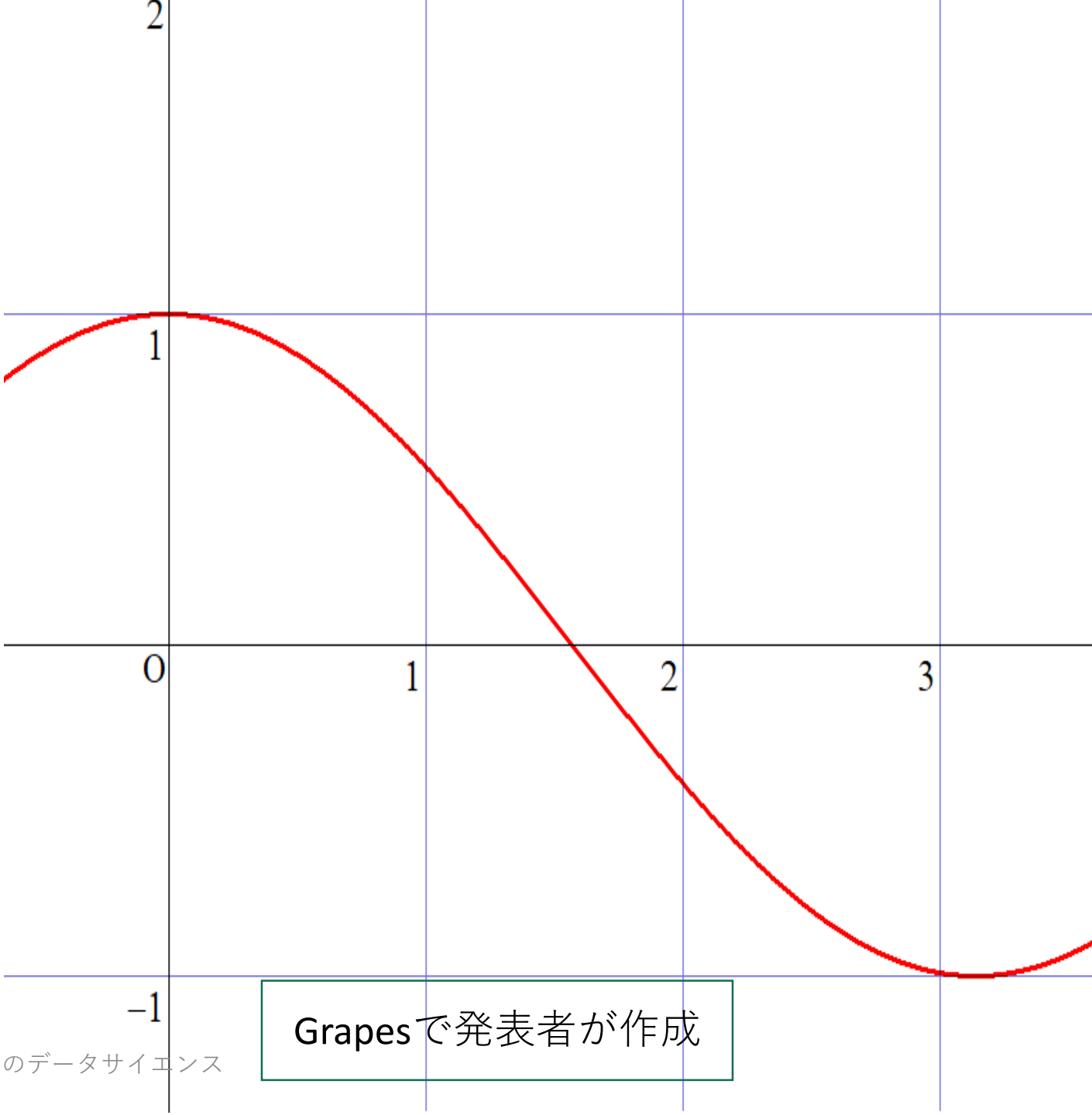
相関係数をベクトルの内積で解釈（同値）

•

$$\rho_{xy} = \cos\theta = \frac{x \cdot y}{|x||y|} \quad \cdots\cdots (2)$$

相関係数は
順序尺度ではない

- 相関係数 が COSの値なら
  比例的な変化はしない。
- COSの値が0近辺では比例
  に近い。1近辺では、
  非常に鈍感である。



Grapesで発表者が作成

疑問③の解決として
X ,Yは、データなのか確率変数なのか？

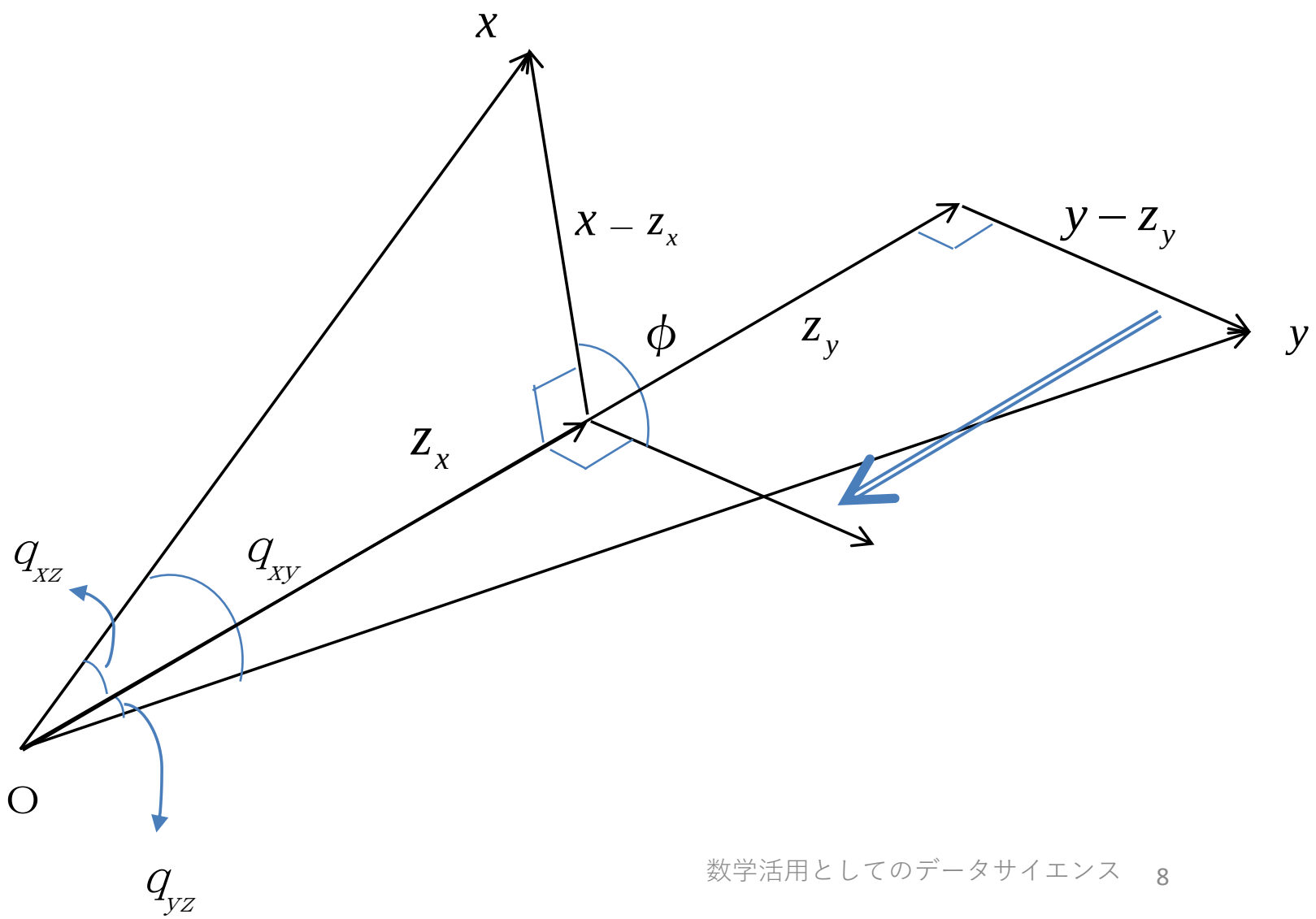
$$\rho_{xy} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \quad \cdots\cdots \quad (1)$$

- データだとすると記述統計としての統計量 そのデータ内の判断にとどめた方が安全。
- 確率変数だとすると背後に2次元正規分布を仮定して、調査データをサンプルとして母集団の相関係数の推定が可能。（検定もできる。しかし、内容を範囲を大きく逸脱）

新しい統計量の開発も不可能ではない

- ベクトルを用いてデータ間の構造を幾何学的にモデル化する。
- 疑似相関を引き起こすベクトルzを取り除いたベクトルで相関係数を考える

→　ベクトルzの影響を受けない相関係数が求められる。



偏相関係数

$$\rho_{xy\cdot z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho^2_{xz}}\sqrt{1-\rho^2_{yz}}} \quad \cdots\cdots \quad (3)$$

偏相関係数は先ほどの幾何学モデル解釈と同値

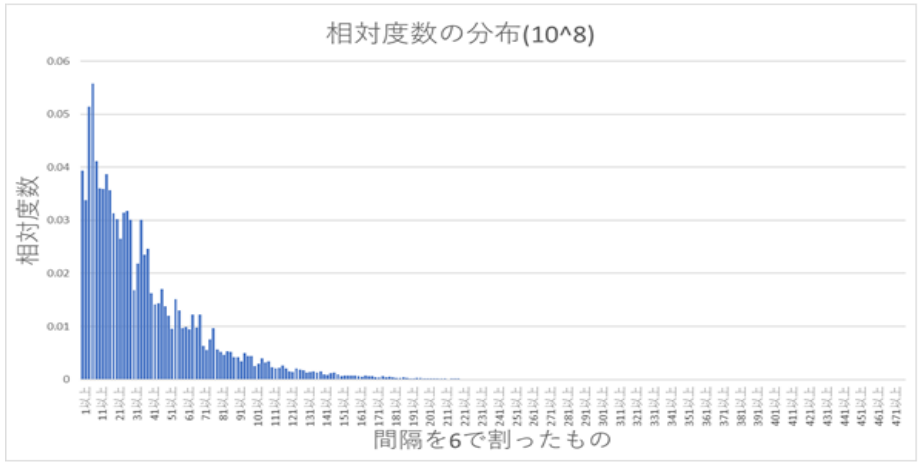
（詳細は、2019年度 附属高校の研究紀要をご参照ください）

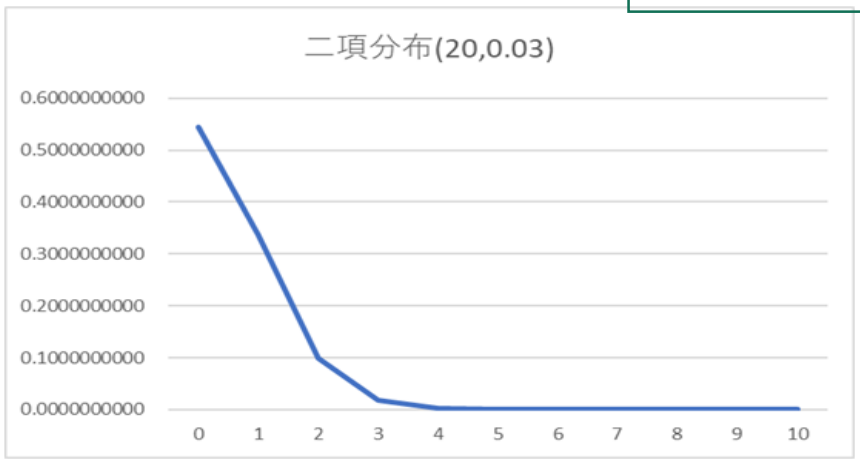
$$\rho_{xy\cdot z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1-\rho^2_{xz}}\sqrt{1-\rho^2_{yz}}} = \frac{(x-z_x)\cdot(x-z_y)}{|x-z_x||x-z_y|} = \cos\phi$$

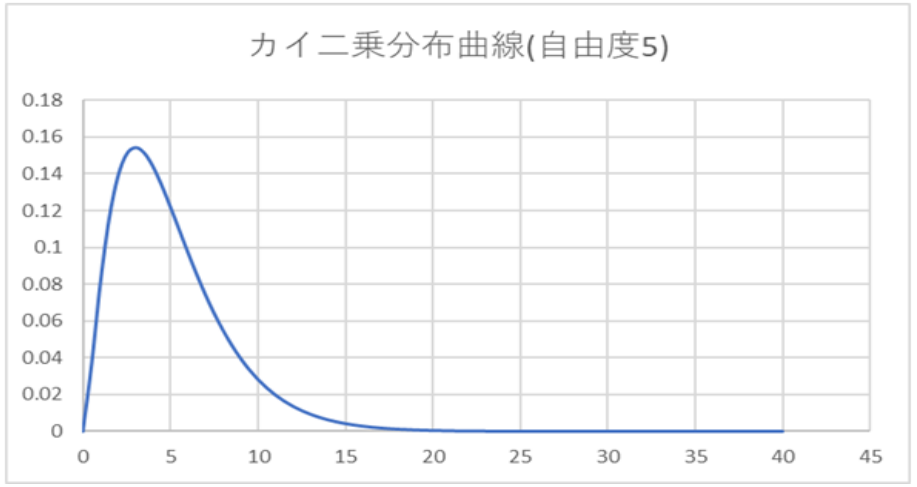
数学の問題にデータサイエンスを

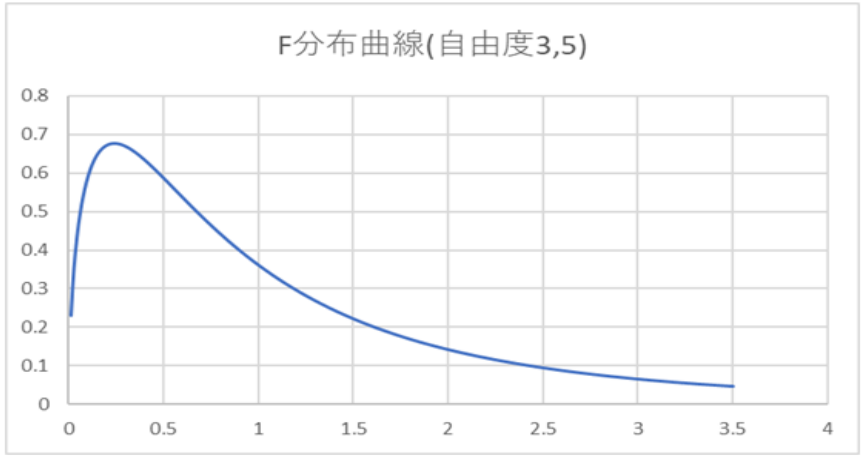
双子素数問題にチャレンジ　距離のヒストグラム

Excelで生徒が作成





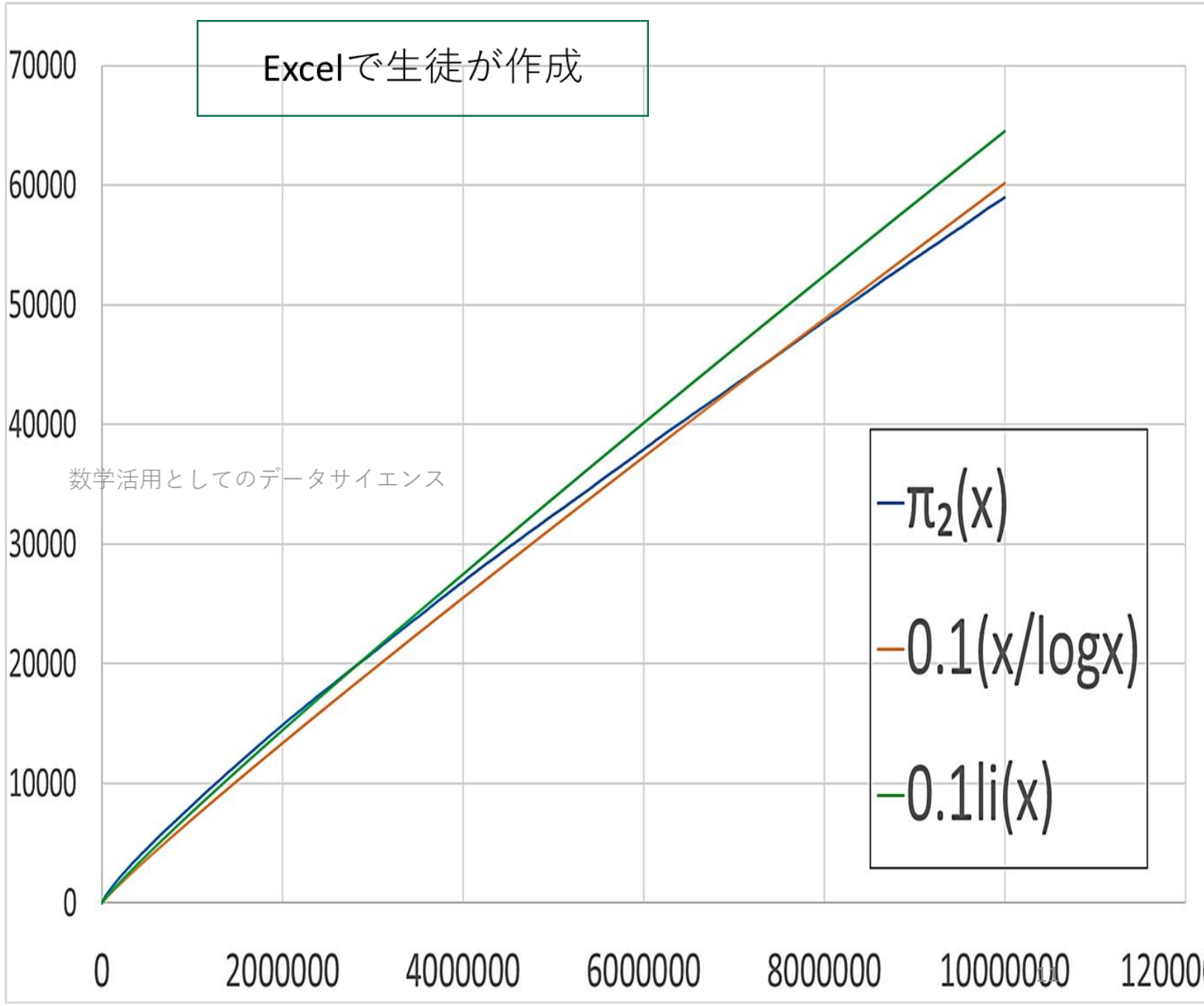




素数定理に関係する関数とグラフを比較（生徒作成Excelによる）

双子素数問題は、素数定理のみでは解決できないのでは？

数学の難問にデータサイエンス的アプローチをすることで、難問の理解・や解決の予想につながる。



＜まとめ＞
データサイエンスと数学を相互活用すると

- データサイエンスに数学を活用し、数学にデータサイエンスを活用する。 個別に教えるよりも、活用させあうことで双方の理解が深まる。
- 科学的に探究する姿勢・態度が育成される。
- 授業開発、教材開発の可能性が多くある。

※コメントを付していない　資料中の数式、図形などはWordによって発表者が作成、または、作成した図形を貼り付けたものである。