教育用標準データセット（SSDSE）の公開とそれに基づく統計教育の推進

（独）統計センター
椿　広計

2018・11・16

官民オープンデータの利活用の動向及び人材育成の取組

現状の把握：データの価格破壊

- 2018年6月15日閣議決定：未来投資戦略2018
- ―「Society 5.0」，「**データ駆動型社会**への変革」―
  - AI,ビッグデータ，IoTの社会実装がもたらしたもの
  - Industry 4.0と呼ばれる産業構造の変革が世界的に進行
- ビッグデータ時代？⇒**チープデータ時代**
  - 「限界費用ゼロ社会」
    - Rifkin, J. (2014) *The Zero Marginal Cost Society – The Internet of Things and the Rise of the Sharing Economy*, St Martins Pr.
    - 柴田裕之訳 (2015) 限界費用ゼロ社会<モノのインターネット>と共有型経済の台頭, NHK出版.

データサイエンス教育の重要性と設計

- データの価値＝データから派生する知識価値（ニーズ依存）
  ーデータとその処理の価格（負の経済価値）
  - **コスト低減⇒データ駆動サービス産業革命**
- 効率的・効果的知識価値創成の横断科学～計測と統計
- 育成すべき人材像と教育のデザイン
  - **学生の活動，教授・教員の活動，教程，教材，教育方法**
  - 最終目標活動：創造的価値創生アクティブラーニング～自由課題
    - 目的を持った創造的情報収集と分析
  - 前段としての**共通課題**とその教材：
    - 自由度の高い「標準データ」とその自律的かる多様な分析の生徒間での共有
  - **標準データセットの要求品質**
    - 親近性「自分たちのデータ」：学生や教員になじみがあること
    - 具体性：個別データについても意味が分かり議論できる
    - 多様性：様々な課題抽出が可能

2018/06/27 統計局・統計センターSSDSE
Standardized Statistical Data Set for Education
https://www.nstac.go.jp/SSDSE/

- **都道府県・市町村のすがた（社会・人口統計体系）**
  - 全市区町村（東京23区を含む）：1741自治体
    - 791市, 744町，183村，23区
    - 生徒誰もがどこかに所属．有名な市町村も多い．
  - 111変量抽出：完備データセットに加工
    - 人口・世帯，自然環境，経済基盤，行政基盤，教育
    - 文化・スポーツ，居住，健康・医療，福祉・社会保障
  - 大学データサイエンス教育，Good Practice **共有化**への活用
- 教育用標準データの量的階層性
  - 都道府県データ・教室内データ取得（50件規模）：小中学校
  - 市町村データ，上場企業データ（1500～2500件規模）：高校・大学
  - 定型的ビッグデータ（数万件規模）：匿名ミクロデータ：大学院
  - 画像・音声などのビッグデータ

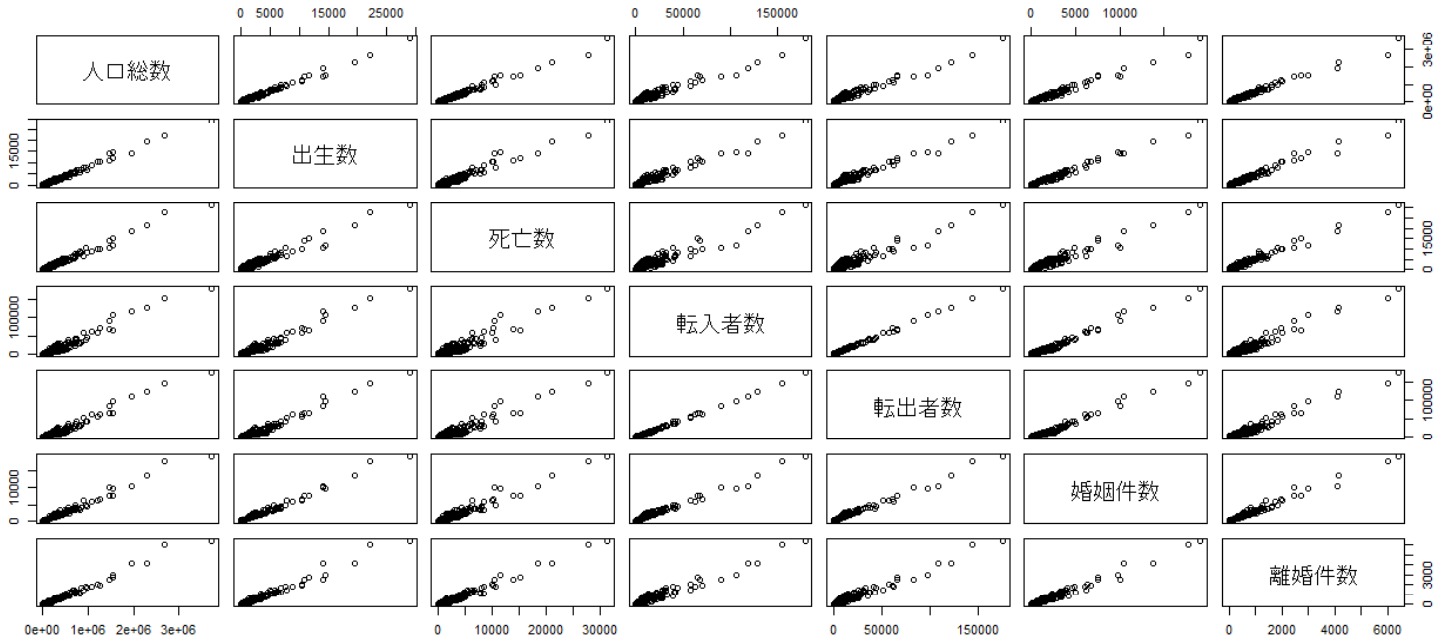
変数一覧：人口・世帯・事業所

[0]人口総数　[1] "人口総数男"　"人口総数女"　"日本人人口"　"日本人人口男"
[5] "日本人人口女"　"１５歳未満人口"　"１５歳未満人口男"　"１５歳未満人口女"
[9] "１５から６４歳人口"　"１５から６４歳人口男"　"１５から６４歳人口女"　"６５歳以上人口"
[13] "６５歳以上人口男"　"６５歳以上人口女"　"７５歳以上人口"　"７５歳以上人口男"
[17] "７５歳以上人口女"　"外国人人口"　"出生数"　"死亡数"
[21] "転入者数"　"転出者数"　"世帯数"　"一般世帯数"
[25] "一般世帯人員数"　"核家族世帯数"　"単独世帯数"　"６５歳以上の世帯員のいる核家族世帯数"
[29] "高齢夫婦のみの世帯数"　"高齢単身世帯数６５歳以上の者１人"　"婚姻件数"　"離婚件数"
[33] "総面積北方地域及び竹島を除く"　"可住地面積"　"事業所総数"　"事業所数農業林業"
[37] "事業所数建設業"　"事業所数製造業"　"事業所数電気ガス熱供給水道業"　"事業所数情報通信業"
[41] "事業所数運輸業郵便業"　"事業所数卸売業小売業"　"事業所数金融業保険業"　"事業所数不動産業物品賃貸業"
[45] "事業所数学術研究専門技術サービス業" "事業所数宿泊業飲食サービス業"　"事業所数生活関連サービス業娯楽業"
"事業所数教育学習支援業"
[49] "事業所数医療福祉"　"事業所数複合サービス事業"　"事業所数サービス業他に分類されないもの"
"事業所数公務他に分類されるものを除く"

変数一覧続き：事業所，従業員，財政，施設

[53] "第１次産業事業所数"　"第２次産業事業所数"　"第３次産業事業所数"　"従業者総数"
[57] "従業者数農業林業"　"従業者数建設業"　"従業者数製造業"　"従業者数電気ガス熱供給水道業"
[61] "従業者数情報通信業"　"従業者数運輸業郵便業"　"従業者数卸売業小売業"　"従業者数金融業保険業"
[65] "従業者数不動産業物品賃貸業"　"従業者数学術研究専門技術サービス業"
"従業者数宿泊業飲食サービス業"　"従業者数生活関連サービス業娯楽業"
[69] "従業者数教育学習支援業"　"従業者数医療福祉"　"従業者数複合サービス事業"
"従業者数サービス業他に分類されないもの"
[73] "従業者数公務他に分類されるものを除く"　"第１次産業従業者数"　"第２次産業従業者数"　"第３次産業従業者数"
[77] "経常収支比率市町村財政"　"実質公債費比率市町村財政"　"歳入決算総額市町村財政"　"地方税市町村財政"
[81] "歳出決算総額市町村財政"　"民生費市町村財政"　"土木費市町村財政"　"教育費市町村財政"
[85] "災害復旧費市町村財政"　"幼稚園数"　"幼稚園在園者数"　"小学校数"
[89] "小学校教員数"　"小学校児童数"　"中学校数"　"中学校教員数"
[93] "中学校生徒数"　"高等学校数"　"高等学校生徒数"　"公民館数"
[97] "図書館数"　"総人口非水洗化人口.水洗化人口"　"非水洗化人口"　"小売店数"
[101] "飲食店数"　"大型小売店数"　"一般病院数"　"一般診療所数"
[105] "歯科診療所数"　"医師数"　"歯科医師数"　"薬剤師数"
[109] "保育所等数"　"保育所等在所児数"

例えば人口関係の統計について，
散布図行列を描くと次のようになります



	Factor1	Factor2	Factor3
１５歳未満人口		−0.195	−0.658
１５から６４歳人口	0.135	−0.233	−0.731
外国人人口	0.161		−0.263
核家族世帯数		−0.125	0.417
単独世帯数	0.242		0.298
６５歳以上の世帯員のいる核家族世帯数			0.958
高齢夫婦のみの世帯数			0.965
高齢単身世帯数６５歳以上の者１人			0.849
従業者数農業林業			0.285
従業者数建設業	0.441		0.158
従業者数製造業	0.326		−0.283
従業者数電気ガス熱供給水道業	0.305		
従業者数情報通信業	0.972		
従業者数運輸業郵便業	0.373		−0.169
従業者数卸売業小売業	0.955		
従業者数金融業保険業	0.965		
従業者数不動産業物品賃貸業	0.982		
従業者数学術研究専門技術サービス業	0.866		
従業者数宿泊業飲食サービス業	0.656		
従業者数生活関連サービス業娯楽業	0.657		
従業者数教育学習支援業	0.695		
従業者数医療福祉	0.520		0.179

自治体特性の測定モデル：3因子モデル（因子分析：変数は絞り込んでいます）
サービス化，文化教育化，高齢化

従業者数複合サービス事業	0.210		0.363
従業者数サービス業他に分類されないもの	0.977		
従業者数公務他に分類されるものを除く	0.834		0.147
実質公債費比率市町村財政		0.960	
地方税市町村財政		0.956	
教育費市町村財政		0.974	
災害復旧費市町村財政		0.922	
中学校数		0.880	
公民館数		0.707	0.141
図書館数			
非水洗化人口		0.868	
小売店数	0.359		0.468
飲食店数	0.660		
一般診療所数	0.245	0.580	
歯科診療所数	0.667		
医師数	0.489		
歯科医師数	0.886		
薬剤師数	0.911		
保育所等数		0.700	
保育所等在所児数		0.145	−0.224

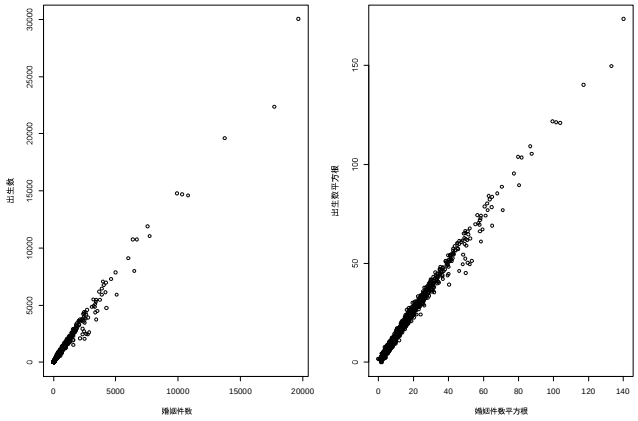
一見推測統計風の記述統計

- 与えられたデータへの統計モデル当てはめ
- 残差分析で系統変動が残らないことを保証
- 匿名化されていないデータだからこそ気づく残差の系統変動

	都道府県	市区町村	残差
1113	大阪府	大阪市	-17.5
643	東京都	豊島区	-17.3
631	東京都	新宿区	-16.1
641	東京都	中野区	-15.8
637	東京都	目黒区	-13.9
642	東京都	杉並区	-12.7
638	東京都	大田区	-12.5
636	東京都	品川区	-12.2

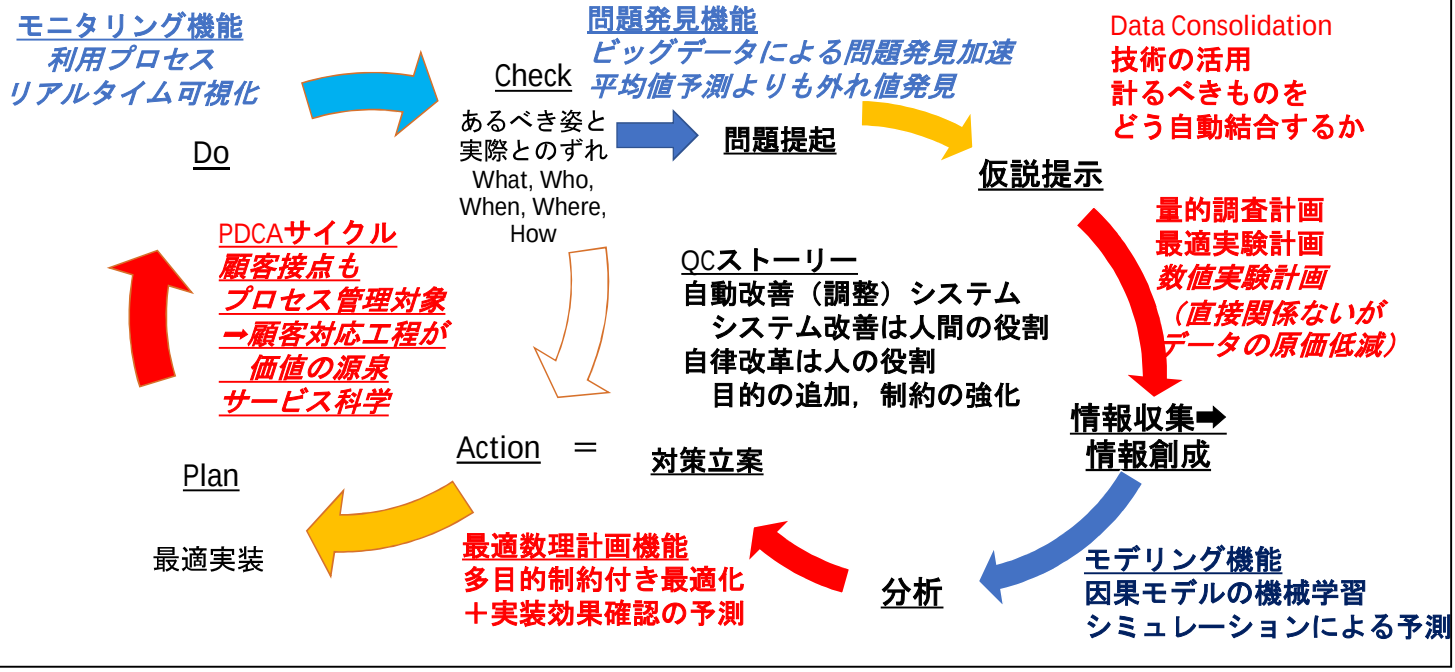
婚姻件数と出生件数との散布図
出生件数∝婚姻件数！

原データ　　両平方根変換データ



等分散性の仮定が成立しそうなのはどちら？

デミング・石川モデルのSHINKAの方向性
データサイエンスによる問題解決

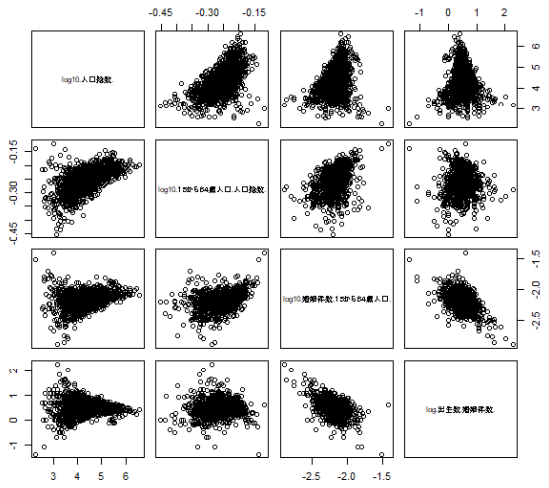


統計的問題とその分解

- 統計的問題
  - 入力が不適切なので狙い値からずれる
    - 入出力の関係を使って出力を調整
  - そもそもデータがばらついていて，狙いの範囲に入らない
    - バラつきの原因を探る
- 問題の分解
  - 利益が少ない
    - 利益＝売上ーコスト
      - 売り上げが少ない
      - コストがかかる
  - 出生数が少ない
    - 出生数
      ＝人口×
      （出生可能年齢者数/人口）×
      婚姻件数/出生可能年齢者数×
      出生数/婚姻件数

      高齢化，未婚化，少子化

人口,婚姻件数，出生数>0の1728自治体
問題分解後の両対数プロットと相関
何が困難な問題でしょうか？



相関係数 全て対数変換	人口	15-64歳人口 /人口	婚姻件数 /15-64歳 人口	出生数 /婚姻件数
人口	1.00	0.58	0.30	0.02
15-64歳人口/人口	0.58	1.00	0.39	0.03
婚姻件数/ 15-64歳人口	0.30	0.39	1.00	-0.49
出生数/ 婚姻件数	0.02	0.03	-0.49	1.00

(婚姻件数／１５から64歳人口)の平方根と都市類型の関係性

- Residuals:比較的よく説明できている

```
      Min        1Q    Median        3Q       Max
-0.054144 -0.005226  0.000303  0.005549  0.094634
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.0839855  0.0002538 330.883   <2e-16 ***
Factor1          0.0165198  0.0010645  15.519   <2e-16 ***
Factor2          0.0112804  0.0009674  11.661   <2e-16 ***
Factor3         -0.0031938  0.0002561 -12.472   <2e-16 ***
Factor1:Factor2 -0.0189579  0.0013812 -13.725   <2e-16 ***
Factor1:Factor3  0.0062619  0.0006952   9.007   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05’
1Residual standard error: 0.01055 on 1724 degrees of freedom
Multiple R-squared:  0.7157,    Adjusted R-squared:  0.7148
F-statistic: 867.8 on 5 and 1724 DF,  p-value: < 2.2e-16
```

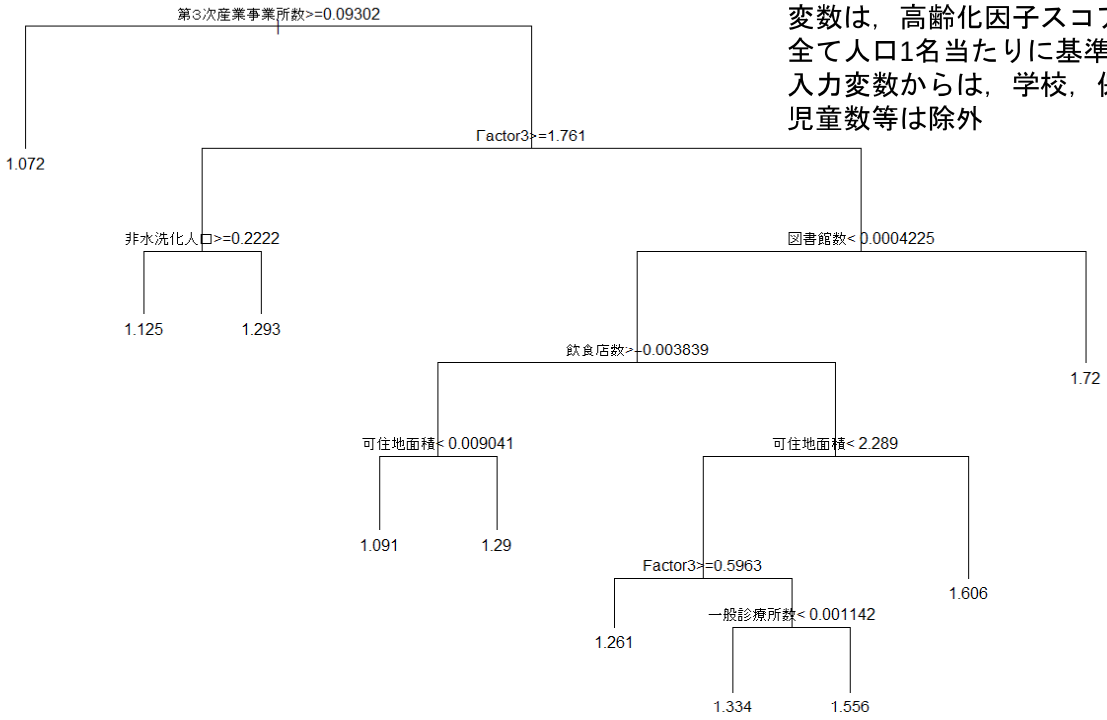
実測値＞モデル予測値
一番ずれているのは,
（予想より婚姻件数が大きい）
残差MAX:0.09
福島県小野町
　人口：10475人
15~64歳人口：5991人
　婚姻件数：29,出生数：60
　小野町結婚世話焼き人　16名

予測が外れることは
発見のチャンス

(出生件数／婚姻件数)の平方根と都市類型の関係性

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.83590 -0.08620 -0.00931  0.07036  1.88779
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.296084   0.004659 278.212  < 2e-16 ***
Factor1                -0.130891   0.022059  -5.934 3.58e-09 ***
Factor2                -0.010154   0.028041  -0.362  0.71730
Factor3                -0.024888   0.004463  -5.577 2.84e-08 ***
Factor1:Factor2         0.205302   0.039877   5.148 2.93e-07 ***
Factor1:Factor3        -0.027014   0.015434  -1.750  0.08025 .
Factor2:Factor3         0.019060   0.018467   1.032  0.30215
Factor1:Factor2:Factor3 -0.090928   0.034396  -2.644  0.00828 **
Residual standard error: 0.1809 on 1722 degrees of freedom
Multiple R-squared:  0.04468,    Adjusted R-squared:  0.0408 ⇒説明力が弱い
F-statistic: 11.51 on 7 and 1722 DF,  p-value: 2.392e-14
```

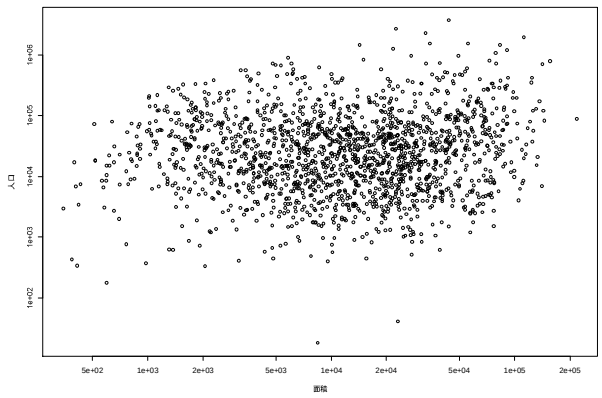
(出生数/婚姻件数)の平方根の回帰の樹：CART （Breiman et al., 1984)
第2世代人工知能（自動層別：Black Box型ではない）



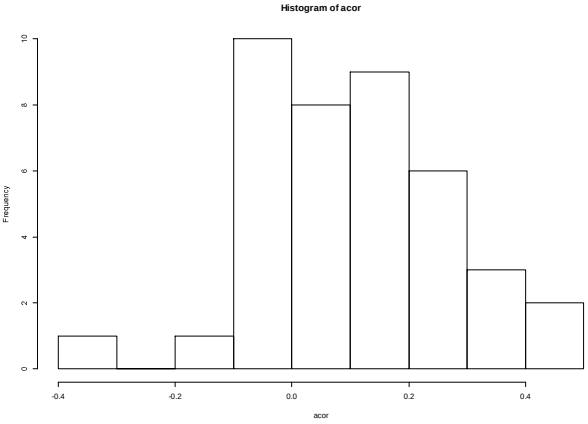
変数は，高齢化因子スコアを除いて
全て人口1名当たりに基準化
入力変数からは，学校，保育園，
児童数等は除外

推測統計の勉強にも使える？
標本の大きさ50の相関係数の分布；厳密な分布は難しくても

自治体面積と人口の相関
散布図



無作為標本の大きさ50で相関係数を
40回計算した場合の標本分布



SSDSEによる
Good Modelling Practicesの知の全国共有

- **第1回統計データ分析コンペティション**
  - **次代を担う高校生・大学生等の統計リテラシーを向上**
  - 主催：総務省，(独)統計センター，(一社)日本統計協会
  - 後援：国立研究開発法人科学技術振興機構，
    一般社団法人日本統計学会，全国統計教育研究協議会
  - 受賞論文の決定・発表：2018年10月18日（統計の日）
    - 高校生の部総務大臣賞
      - 大段利々子(広島大学附属高等学校)
        - 本当に日本の医療は危機的状況にあるのか？
    - 大学生・一般の部総務大臣賞：私の分析例よりずっと力作ぞろいです！
      - 平原幸輝（早稲田大学人間学部人間環境学科）
        - 地方創生における三つの「鍵」

ぜひ，皆さんもSSDSEで色々なデータ分析教育を始めて下さい

統計センターはこれからもSSDSEのバージョンアップに努めますので，
様々なご意見を頂戴できれば幸いです