



データをサイエンスする

統計数理研究所・統計思考院 特任教授
(独)統計センター 特別研究員
中央大学工学部 客員教授
田村義保

数理・データサイエンス教育強化拠点コンソーシアム
2020年度 関東・首都圏ブロック 第8回ワークショップ
～農学分野における数理・データサイエンス・AIの教育、研究
の展開普及～
2021年3月25日
オンライン





自己紹介(0-1)



<https://www.hpcwire.jp/archives/20001>



自己紹介(0-2)

田村義保氏 略歴



1975年 東京工業大学理学部物理学科卒業
 1980年 東京工業大学大学院理工学研究科物理学専攻博士課程修了（理学博士）
 1980年 日本学術振興会奨励研究員
 1981年 統計数理研究所研究員
 1985年 統計数理研究所助手
 1986年 統計数理研究所助教授・総合研究大学院大学助教授（兼任1989年）
 1997年 統計数理研究所教授・総合研究大学院大学教授（兼任）
 1998年 統計数理研究所・技術課長(2000年まで)
 2000年 統計数理研究所・統計計算開発センター長(2004年まで)
 2004年 統計数理研究所・副所長(2010年まで)
 2010年 統計数理研究所・データ科学研究系主幹(2011年まで)
 2011年 統計数理研究所・副所長(2017年まで)
 2017年 統計数理研究所・データ科学研究系主幹(2018年まで)
 2018年 統計数理研究所 名誉教授・特任教授・総合研究大学院大学 名誉教授
 2018年 独立行政法人統計センター 統計技術研究課 特別研究員

<https://www.hpcwire.jp/archives/20001>



自己紹介(0-6)

ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)

【わがパソコン人生】第10回 田村義典 X +

← → ↺ ↻ 🔒 https://www.hpcwire.jp/archives/20001 🔍 検索

よく見るページ Firefoxを使いこなそう A C S U(選手) Y 桜並木でお花見RUN あ... 他ブックマーク

ビッグデータを扱っていました。

—長年データサイエンスを研究されてきた田村先生は、今のAI時代をどう見ているのでしょうか？

統計学者の中には統計学が上にあるとあって、下にデータサイエンスがあるという人もいます。私はデータサイエンスが上にあり、そのひとつがAIで、ひとつが統計学という手法だと思っています。だから、統計的手法が大事なのではなく、AI的な情報学的手法も大事で、両方をうまく使えれば良いのですが、AIの人で良くないのは、それはどう考えても重回帰分析だろうという統計学上の分析をしておきながら、AIを使いましたというケースです。マスコミがAIと言ったほうがウケるから言っているのか、本人がそう言っているのかは分かりませんが、ブラックボックス的な解析はすべてAIになってしまっています。

字種の分野はなぜだか、どこまでがAIでどこまでが統計が分けたがる人がいますが、私は分ける必要はないと思っています。正直なところ、どこまでがAI的な手法で、どこまでが統計的な手法となるのか、その分類方法も分かりませんし、気にすることもないと思っています。

今は何でもビッグデータを使ってやりましたと言うけれど、それもおかしいだろうと思います。売上データを使いましたか、もう少し細かく言った方が良いでしょう。ビッグデータもどれぐらいがビッグデータか。そもそも、ビッグデータの定義も実は総務省の定義と経産省の定義が微妙に違うんですよ。私が個人的にビッグデータで一番良い定義だと思ったのは、某先生が言われた「Excelで扱えない量がビッグデータ」です。メモリが多くてもExcelでは100万行くらいまでしか読めませんから。

今はスパコンもデータサイエンス、データ分析で使うような風潮になっていますが、助教になったあと、計算機担当として多くの業者の方と会い、カタログもたくさんもらいましたが、彼らのカタログに記載されていたスパコンの使う目的は流体計算といったもので、「データ解析」「統計解析」とは書いてありませんでした。そこで、業者の方々に「データ解析」や「統計解析」と書くよう、1987,8年ごろから提案をし続けていますが、未だに書いてもらっていません。書いてあれば、それなりに皆が注目したと思うのですが。

—統計研のような研究所は欧米にもあるのでしょうか？

同じような研究所は台湾とインドにしかないのではと思います。台湾、インド、日本は三機関で連携を結び、毎年ワークショップや研究会を持ち回りで行っています。アメリカは主に統計をやっている人が集まっているような研究所はありますが、大学の学部とその付属の研究所といった感じですね。

アメリカは規模が大きい大学には生物統計学部や統計学部がありますが、日本には専門的に統計を教える学部はありません。しかし、文科省がデータサイエンス教育を大事だと考え、北大、東大、京大、阪大、滋賀大、九大、この6大学を統計教育の拠点校として、きちんと学部教育をしようとしています。理系修士入学レベルを1年間に5万人、統計研でも理研 革新知能統合研究(AIP)センターと共同で、業界を代表するような高度なデータサイエンティストを年間50人育てるための活動を行っています。

—田村先生にとってスパコンとは？

研究に欠かせない道具ということですね。それ以上でも、それ以下でもないという感じでしょうか。データ解析もやっていますが、私たちの主な目的と業務はデータ解析のための統計的手法を考えてプログラムにする、動くようにすることだと考えています。各分野の人のために新しいデータ解析手法を作るためには、やはり大きな解析をしないとダメなので、スパコンレベルの計算機がないとプログラムを作ることができません。小さいプログラムを作って、大きくすれば動くでしょという世界ではありません。





データサイエンスとは？

インタビューの文字
起こしなので若干の
言い過ぎもあります。

ビッグデータとAI時代をどう見るか

—長年データサイエンスを研究されてきた田村先生は、今のAI時代をどう見ているのでしょうか？

統計学者の中には統計学が上にあって、下にデータサイエンスがあるという人もいます。私はデータサイエンスが上にあり、そのひとつがAIで、ひとつが統計学という手法だと思っています。だから、統計的手法が大事なのではなく、AI的な情報学的な手法も大事で、両方をうまく使えば良いのですが、AIの人で良くないのは、それはどう考えても重回帰分析だろうという統計学上の分析をしておきながら、AIを使いましたというケースです。マスコミがAIと言ったほうがウケるから言っているのか、本人がそう言っているのかは分かりませんが、ブラックボックス的な解析はすべてAIになってしまっています。



データサイエンスとは？

学問の分野はなぜだか、どこまでがAIでどこまでが統計か分けたがる人がいますが、私は分ける必要はないと思っています。正直なところ、どこまでがAI的な手法で、どこまでが統計的な手法となるのか、その分類方法も分かりませんし、気にすることもないと思っています。

今は何でもビッグデータを使ってやりましたと言うけれど、それもおかしいだろうと思います。売上データを使いましたとか、もう少し細かく言った方がよいのではないのでしょうか。ビッグデータもどれぐらいがビッグデータか。そもそも、ビッグデータの定義も実は総務省の定義と経産省の定義が微妙に違うんですよ。私が個人的にビッグデータで一番良い定義だと思ったのは、某先生が言われた「Excelで扱えない量がビッグデータ」です。メモリが多くてもExcelでは100万行くらいまでしか読めませんから。



データサイエンスとは？

今はスパコンもデータサイエンス、データ分析で使うような風潮になっていますが、助教授になったあと、計算機担当として多くの業者の方と会い、カタログもたくさんもらいましたが、彼らのカタログに記載されていたスパコンの使う目的は流体計算といったもので、「データ解析」「統計解析」とは書いてありませんでした。そこで、業者の方々に「データ解析」や「統計解析」と書くよう、1987,8年ころから提案をし続けていますが、未だに書いてもらっていません。書いてあれば、それなりに皆が注目したと思うのですが。



田村は何をやっているの？

- 物理乱数発生法の研究
- 生命科学領域の研究(2次元形状)
- 生命科学領域の研究(その他)
 - ✓ 非定常時系列解析
 - ✓ 呼吸器系の解析(多変量時系列解析)
 - ✓ 肺がん死亡率予測
 - ✓ 白血病の治療効果
 - ✓ 歯科診断
- 椅子のすわりごち
- 女性の飲料の好み
- 靴底と歩行
- 地震発生 of 記憶に関する研究
- 呼吸リズムに関する研究
- 電力利用予測に関する研究

マルコフ連鎖です。
結構あたっています。

時系列モデルと
ニューラルネットの
予測を比較しています。



考えているコース

農学のための統計入門 記述統計学
グラフ・表による情報集約

1. ヒストグラム
2. 累積相対度数グラフ(累積百分率グラフ)
3. 経験分布グラフ
4. ボックスプロット
5. 円グラフ
6. 帯グラフ
7. 折れ線グラフ
8. チャーノフの顔グラフ
9. 時系列プロット
10. QQ プロット
11. ローレンツ曲線
12. 散布図

1. 度数分布表
2. 相関表
3. 分割表



考えているコース

農学のための統計入門 記述統計学
数値による情報集約

1. 総計値
2. 平均値
3. 最大値・最小値
4. パーセンタイル
5. メディアン(中央値)
6. 第1四分位数
第3四分位数
7. モード
8. 範囲
9. 四分位範囲
10. 分散・標準偏差

11. 平均偏差
12. 四分位偏差
13. 変動係数
14. 四分位変動係数
15. 歪度
16. 尖度
17. 共分散、相関



考えているコース

農学のための統計入門 推測統計学
推測統計学 区間推定 仮説検定 等

1. 確率変数の和の分布
2. 平均値の区間推定
3. 分散の区間推定
4. 比率の区間推定

5. 平均値の検定
6. 分散の検定
7. 平均値の差の検定
8. 分散の比の検定
9. 比率の差の検定

10. 独立性の検定
11. 分散分析
12. 単回帰



考えているコース

農学のための統計入門 多変量解析
多変量解析等

1. 重回帰分析
2. 主成分分析
3. クラスタ分析
4. 判別分析
5. 因子分析
6. 時系列解析

7. 傾向スコア
8. 因果分析
9. 分類のための解析方法
10. カルバック・ライブラーダイバージェンス
11. Fisher情報行列
12. 赤池情報量規準



希望する数学知識

最近の課程の数I、II、A、B

分かっていて欲しい数学知識

微分が最大値を求めるための演算であること。

積分が面積を求める演算であること。

確率密度関数の積分が確率になること。

データは確率変数の実現値であること。



分かっているとうれしい数学知識

ベクトル・行列の知識
行列の積や転置行列の概念

逆行列の意味
主成分の意味



計算機リテラシー

WindowsパソコンかMacが使える。
WindowsとMacで文字コードや改行コードが違うことを理解している。
Rでのデータの読み込み方を知っている。
簡単なRコマンドを知っている。

後はやっている内に覚えることができる。



データサイエンスは 役に立つでしょうか？

役に立ちます。しかし、
使いこなすには時間が
かかります。数字への
思い入れが大事！！



ここ5年くらい耳にする言葉

- ビッグデータ
- アナリティクス
- データサイエンティスト
- データ中心科学
- シンギュラリティ
- AI(機械学習、Deep Learning)
-最強の学問.....
-sexy job.....



藤井 聡太(ふじい そうた、2002年7月19日 -)は、日本の将棋棋士。杉本昌隆門下。棋士番号は307。愛知県瀬戸市出身。名古屋大学教育学部附属高等学校在学中。**2021年1月末中退。**

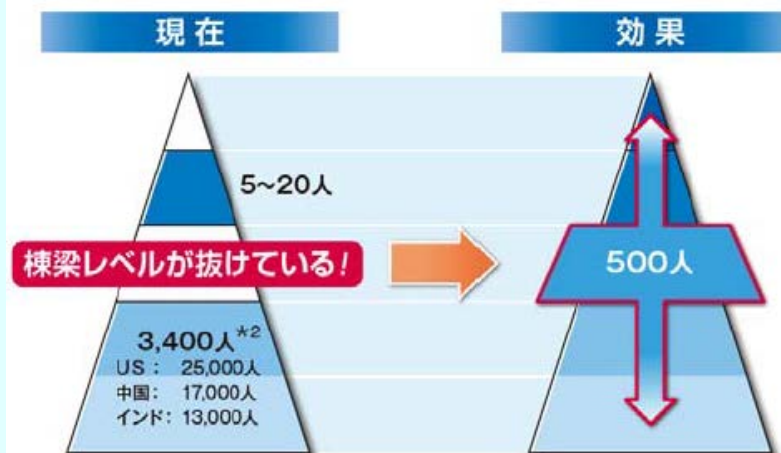
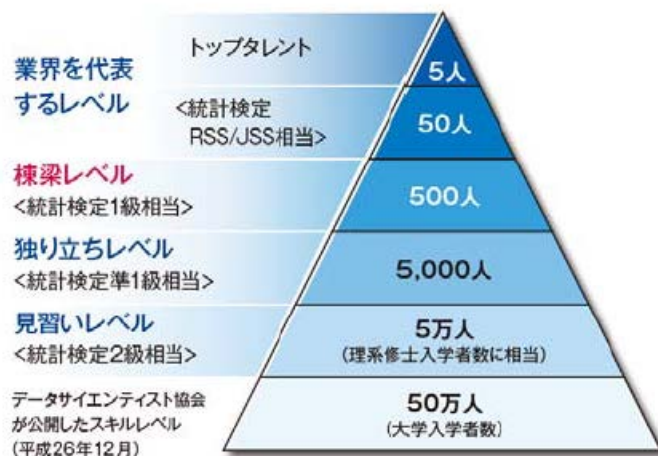
しかし、なんと言っても

◆ 藤井二冠
◆ SDGs
◆ DX



データサイエンティスト

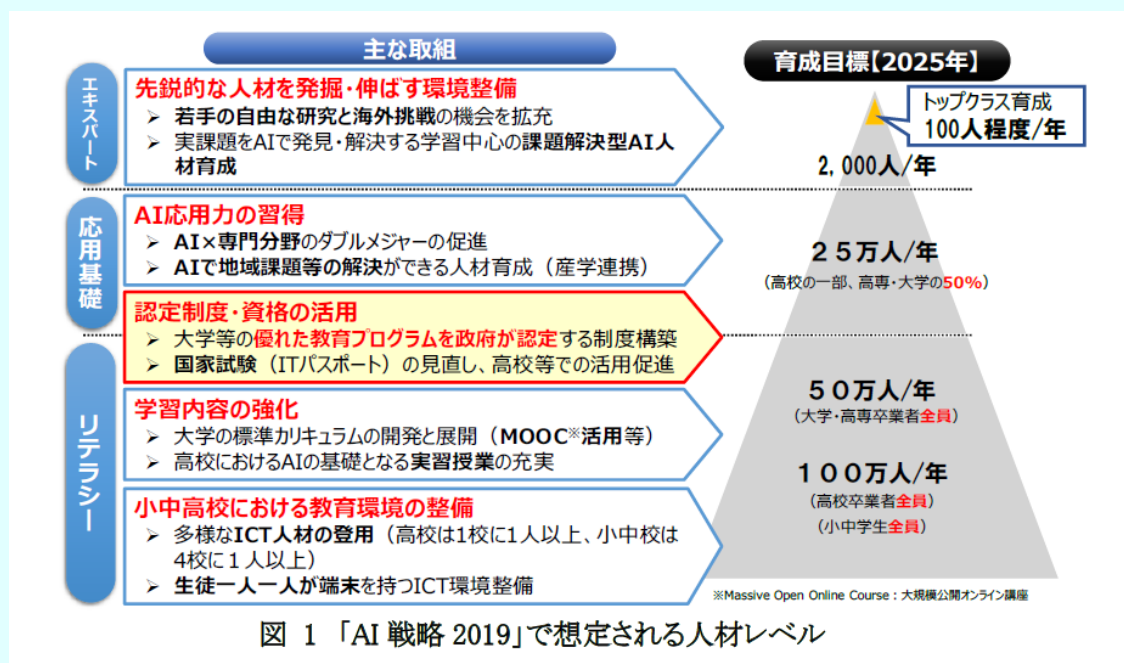
<http://www.ism.ac.jp/shikoin/training/index.html>
から転載



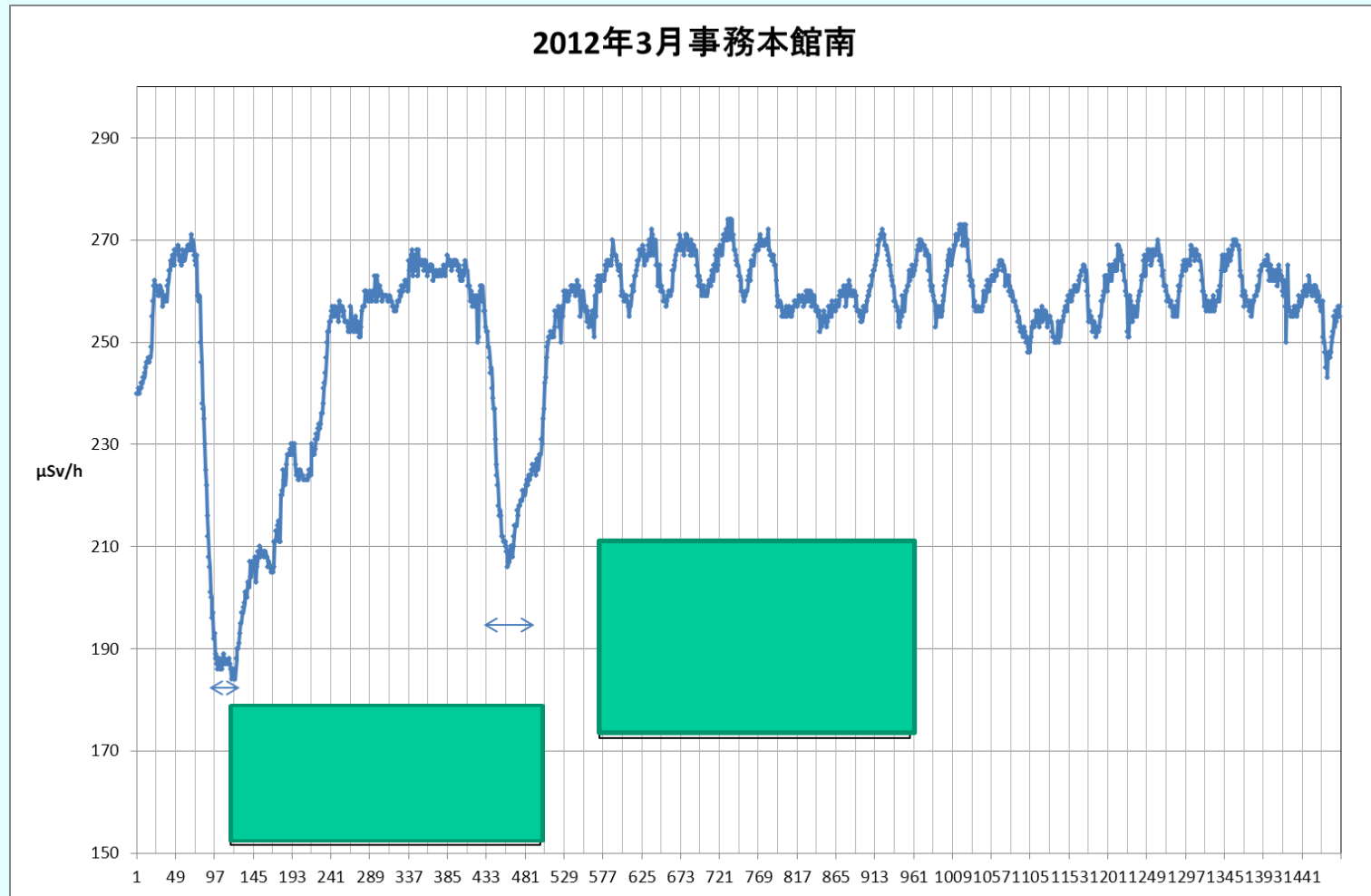


データサイエンティスト

www.kantei.go.jp › [ai_senryaku](#) › [dai6](#) › [siryo1PDF](#)
から転載



記述統計学 はずれ値があったら(1)



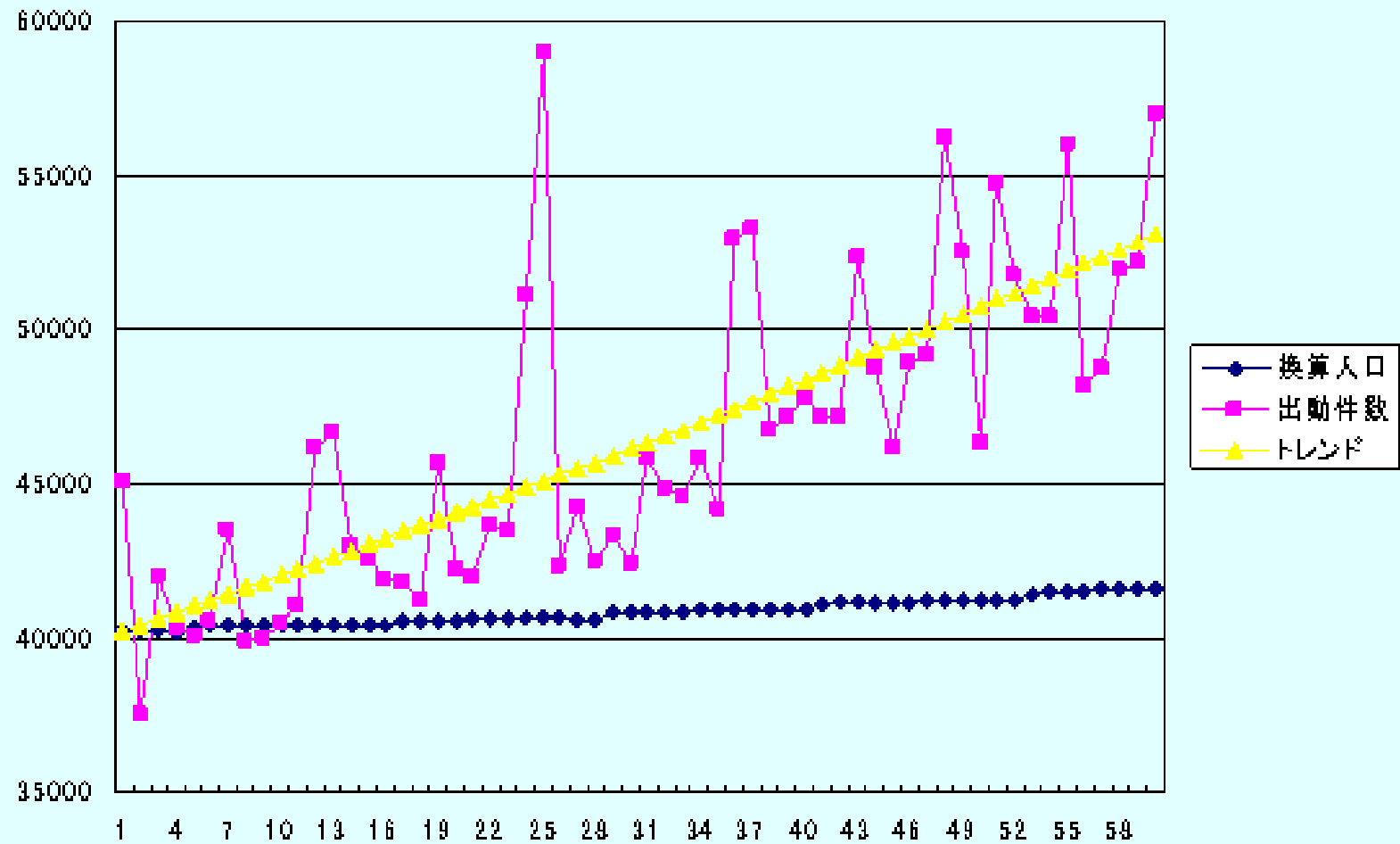


図2.1.4 東京都換算人口と救急出動件数及びそのトレンド