■本資料のご利用にあたって（詳細は「利用条件」をご覧ください）

本資料には、著作権の制限に応じて次のようなマークを付しています。
本資料をご利用する際には、その定めるところに従ってください。

＊：著作権が第三者に帰属する著作物であり、利用にあたっては、この第三者より直接承諾を得る必要があります。

CC：著作権が第三者に帰属する第三者の著作物であるが、クリエイティブ・コモンズのライセンスのもとで利用できます。

🄯：パブリックドメインであり、著作権の制限なく利用できます。

なし：上記のマークが付されていない場合は、著作権が東京大学及び東京大学の教員等に帰属します。無償で、非営利的かつ教育的な目的に限って、次の形で利用することを許諾します。

Ⅰ　複製及び複製物の頒布、譲渡、貸与
Ⅱ　上映
Ⅲ　インターネット配信等の公衆送信
Ⅳ　翻訳、編集、その他の変更
Ⅴ　本資料をもとに作成された二次的著作物についてのⅠからⅣ

ご利用にあたっては、次のどちらかのクレジットを明記してください。

東京大学　Todai OCW　学術俯瞰講義
Copyright 2014, 竹村彰通

The University of Tokyo / Todai OCW　The Global Focus on Knowledge Lecture Series
Copyright 2014, Akimichi Takemura

統計学の今後：
ビッグデータ時代の統計学と
統計教育

工学部計数工学科　竹村彰通

２０１４年６月５日

項目

1. 「統計教育」に関する最近の私の活動
2. 統計的・数理的な分析の例
   - Living histogram の話
   - SNS投稿数データの解析
3. ビッグデータ時代の統計学

1.「統計教育」に関する文章(執筆，編集)

- 数学セミナー2013年5月号

  「統計計算から統計的思考へ／数学教育の中での統計教育の新たな方向」

- 数学セミナー特集 2013年10月号「時代が求める統計学」

  「統計的な考え方と結果の見方」

- 数学セミナー増刊『統計学ガイダンス』2014年7月発行に再録
- 青土社『現代思想』特集2014年6月号「ポスト・ビッグデータと統計学の時代」

  「統計学にとって情報とは何か」

「統計教育」に関する文章(執筆，編集)

- 日本学術会議数理科学委員会数理統計学分科会の提言案の編集 （学術会議で審査中，何らかの形で2014年8月に公表見込）

  「ビッグデータ時代における統計科学教育・研究の推進について」
- 統計検定４級教科書の編集：　「資料の活用」東京図書

- 諸事情により，このような活動をしている
  - 数学と統計学の関係
  - ビッグデータ時代の生き残り
- 研究もしてます!

青土社『現代思想』の目次の一部

【イントロダクション】

ビッグデータと統計学　／　竹内　啓

【討議】

情報（データ）は人を自由にするか　／　西垣　通＋ドミニク・チェン

【インパクト】

統計学にとって情報とは何か　／　竹村彰通

ビッグデータブームを考える　／　水田正弘

ビッグデータは科学を変えたか？　／　出口康夫

【インタビュー】

統計学は科学の文法である　水俣から福島まで、なぜ公害は繰り返されるのか　／　津田敏秀

【統計学の現在】

統計的因果推論の考え方　／　岩崎　学

統計学・確率論の有効性とその限界　／　小島寛之

統計・実証主義・社会学的想像力　／　太郎丸　博

（後略）

2.1　分析の例1
Living histogram の話　(数学セミナー2013/5記事より)

- 人間の身長の分布は正規分布?
- Googleで living histogram を検索!
- 二山の分布に見える. 実は
  - 男女が混ざっているから二山になる
  - 男女を分ける(層別する)と正規分布が見えてくる
- 統計検定4級教科書でもきちんと説明した.

著作権の都合により
ここに挿入されていた画像を削除しました

Mark　F. Schilling, Ann E. Watkins, and William Watkins (2002) Is Human Height Bimodal?, The American Statistician, vol.56 (no.3):223-229, p.223 Fig.1
http://www.tandfonline.com/doi/abs/10.1198/00031300265#.U5-bhCgYuKI

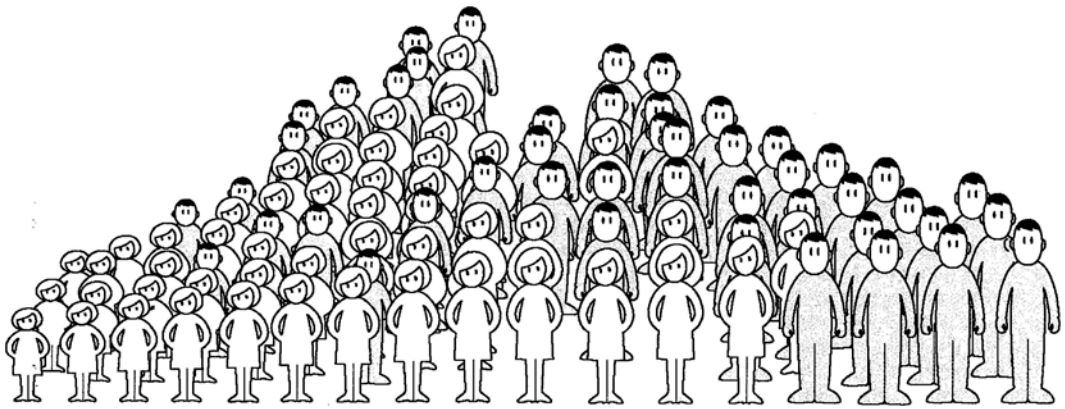
統計検定４級教科書での説明

- 多峰性（峰が複数ある）

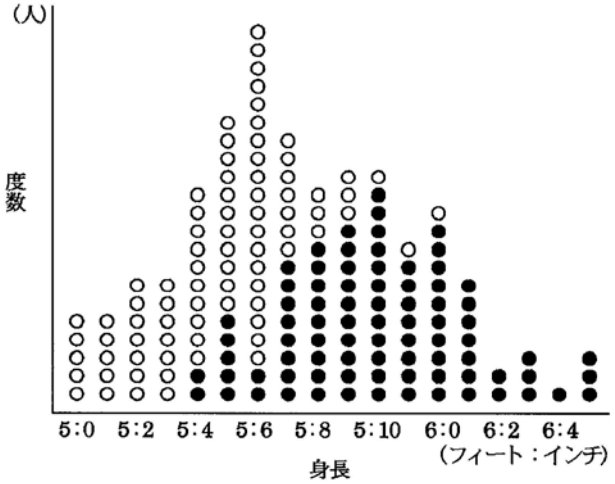
データの集中する部分（ピーク）が2つ以上あるヒストグラムを多峰的といいます．ピークが2つ以上になる場合は，大人と子供などの異質の集団が混在している可能性があります．データを層別（区別）して分析するなどの工夫が必要です．

日本統計学会（編）『日本統計学会公式認定　統計検定4級対応　資料の活用』東京図書、2012年、p.86より引用

数学セミナー編集部による絵



*



*

図版出典：竹村彰通（2013）「統計計算から統計的思考へ」『数学セミナー』、2013年5月号（vol.52, no.5）：50-54、［上図］p.53 図1、［下図］p.53 図2

本当に二山？

- 男女をまぜると身長の分布は本当に二山になるのか？
- Mark F. Schilling, Ann E. Watkins and William Watkins (2002): “Is Human Height Bimodal？”, *The American Statistician*, vol.56, pp.223-229.

  はこの解釈に次のような疑問をなげかけている.
- 正規分布の混合はいつでも二山になるのだろうか.
- 二つの群の平均が十分離れていなければ二山にはならないのではないか.

本当に二山？

- 実はこの疑問にはすでに1904年の Helguerroの論文で答が知られている:

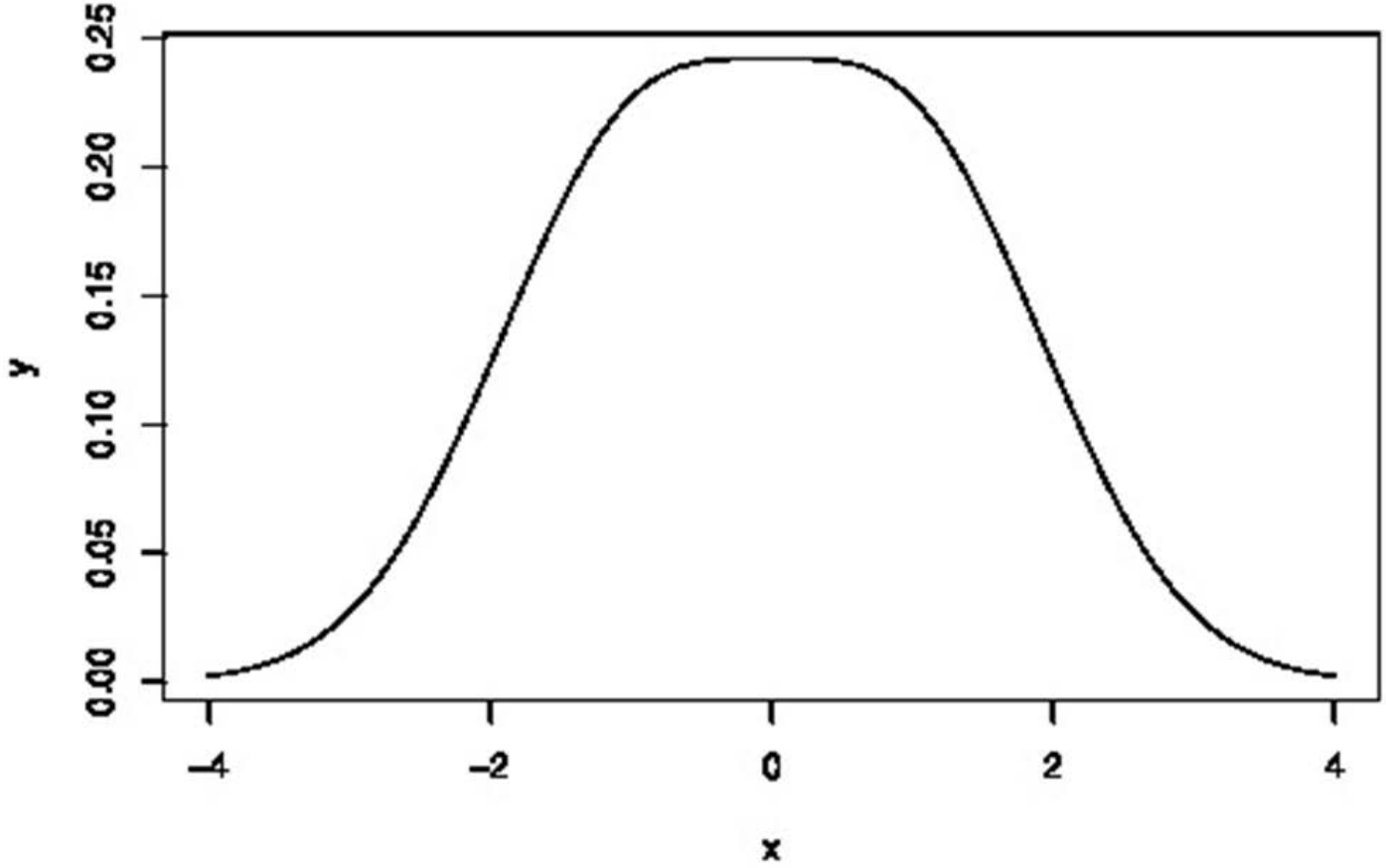
  Helguerro, F. (1904):``Sui Massimi Delle Curve Dimoreche'', Biometrika, vol.3, pp. 85-98.

- 分散 σ^2 を共通とする二つの正規分布 $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ の密度関数を 1/2 の重みで加えた混合分布の密度関数が二山となる必要十分条件は次で与えられる.

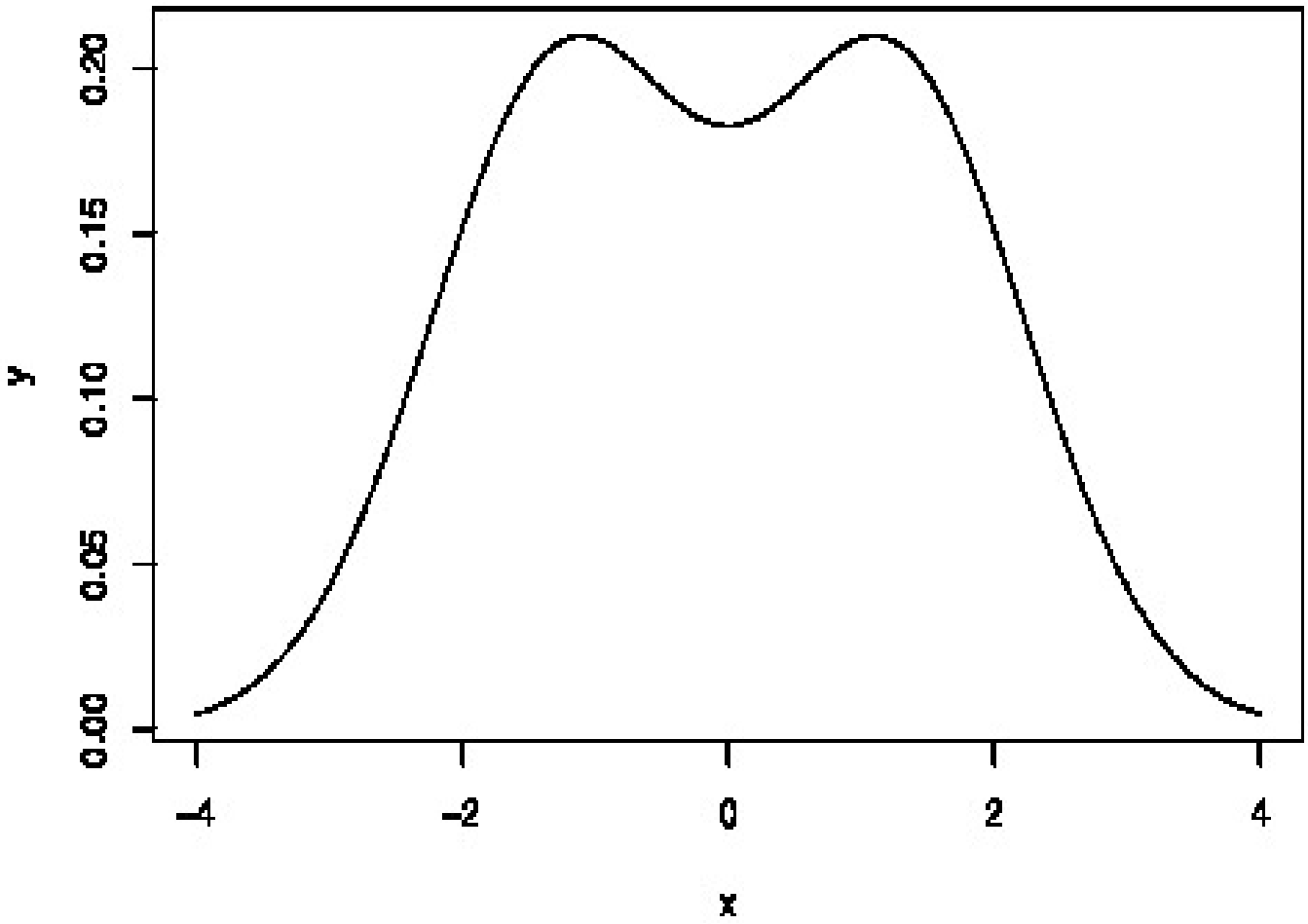
$$|\mu_1 - \mu_2| > 2\sigma$$

- つまり平均同士が標準偏差の2倍離れていないと二山にはならない.
- 平均が丁度標準偏差の2倍離れている時と2.5倍離れている時の図

標準偏差の2倍の場合



標準偏差の2.5倍の場合



本当に二山？

- 男女の平均は標準偏差の2倍も離れているのであろうか？
- Living histogram の写真はひょっとしてやらせではないのか？
- このように現実のデータはなかなか手ごわい…
- Helguerro の結果はすぐにわかるか？

σ=1 としてもよい．でも，次の関数の増減表は結構難しい!

$$f(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_1)^2}{2}\right) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_2)^2}{2}\right) \quad (1)$$

本当に二山？

- 文部科学省の体力・運動能力調査結果を見てみると，20歳の日本人男子の平均身長は約172cm，標準偏差約5.5cm，女子の平均身長は約158cmで標準偏差は男子とほぼ同じである.
- 平均身長の差14cmは標準偏差の2倍の11cmより大きく，密度関数としては二山となる．ただし2.5倍の時の図を見ると，明確にヒストグラムが二山になるほどの差でもない.
- おそらく living histogram の図には「うまく撮影できた」という “publication bias” があるように思われる.
- ちなみに，Helguerro の結果は，正規分布の密度関数が $\mu \pm \sigma$ に変曲点をもっていることを知っていればわかる…

正規分布を教えるのは…

- 高校生に正規分布を教えることは不可能と言われている.
- 正規分布の混合になると，理科系の大学生でも難しい？
- しかし，数学ソフトを使えば，(1)式で $\mu_2 = 0$ とおき，μ_1 を動かしながら $f(x)$ の図をコンピュータで画面表示することは容易.
- コンピュータの使えば，$\exp(-x^2/2)$ の微分ができず，増減表が書けなくても，数学的な事実を(ある程度)納得することができる.
- 一山の分布からの乱数を発生させてヒストグラムをコンピュータに描かせればヒストグラムが二山に見えることも結構多い.

動的グラフィックスの教育的利用の例(立教大学)

2.2 SNS投稿数データの解析
「NTTコム オンライン賞」獲得

第１回データビジネス創造コンテスト
2014年4月19日

投稿数時系列データの
「ベキ減衰モデル」を用いた
消費税増税の消費者への影響の評価

株田達矢，藤山俊文，松井千尋

この分析例の目次

1. 投稿数の典型的パターン
2. ベキ減衰モデル
3. 消費増税の話題性の解析
4. 今後の課題および応用

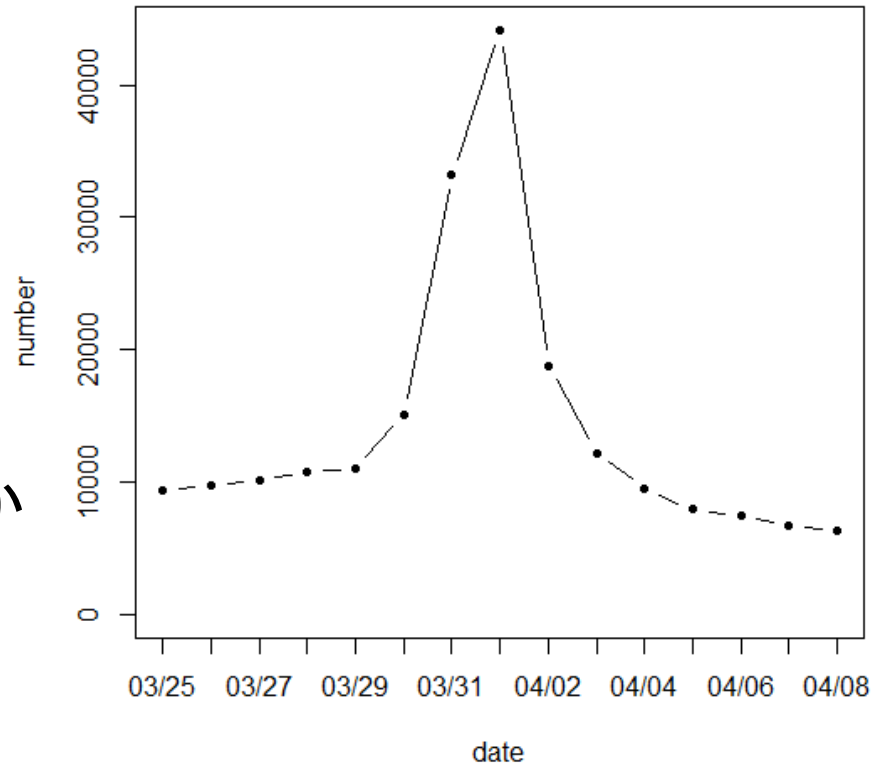
典型的な投稿数のパターン

- 「消費税」を含むブログ投稿件数
- データ提供はNTTコム オンライン・マーケティング・ソリューション株式会社

消費増税前日の3/31と
増税された4/1に
投稿件数は多くなっている

投稿数の増え方・減り方に
何か法則があるのではないか

「消費税」

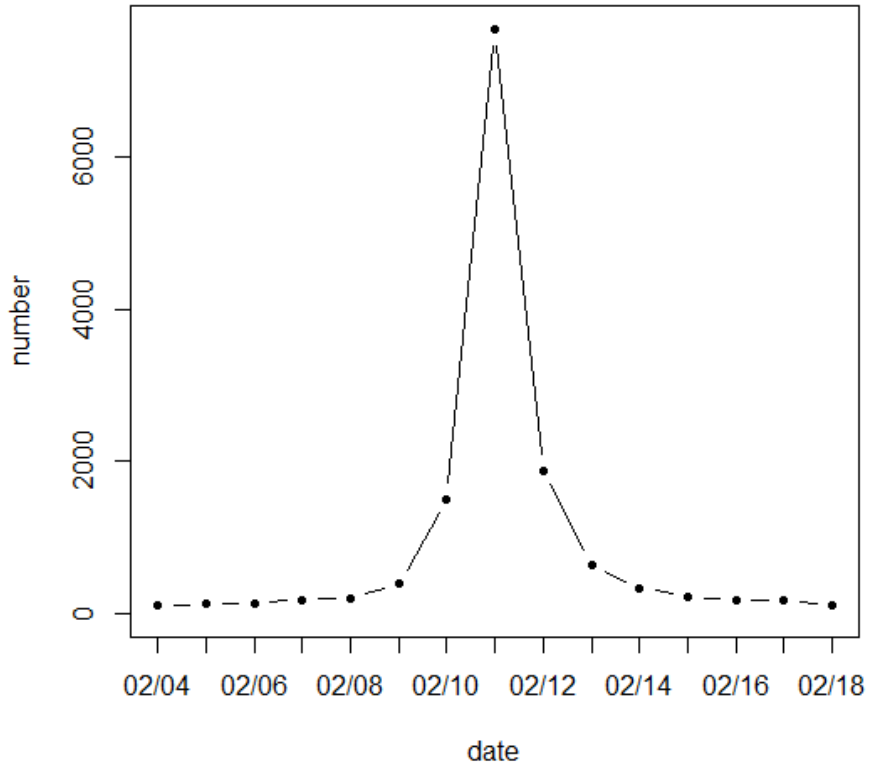


＊画像提供：竹村彰通研究室

典型的な投稿数のパターン

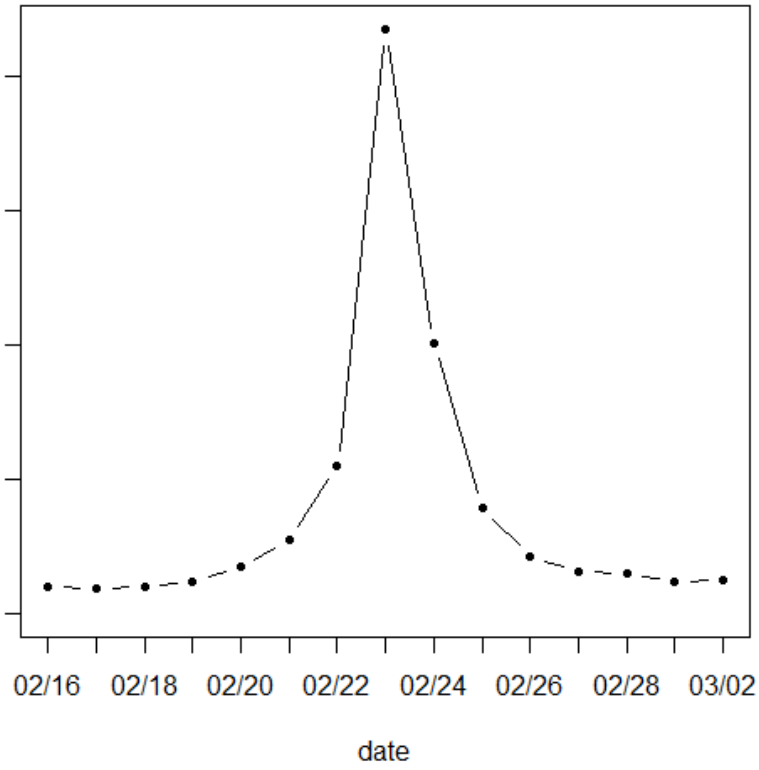
- イベントの前後で投稿数に対称性のある例

2/11「建国記念日」



＊画像提供：竹村彰通研究室

2/23「東京マラソン」



＊画像提供：竹村彰通研究室

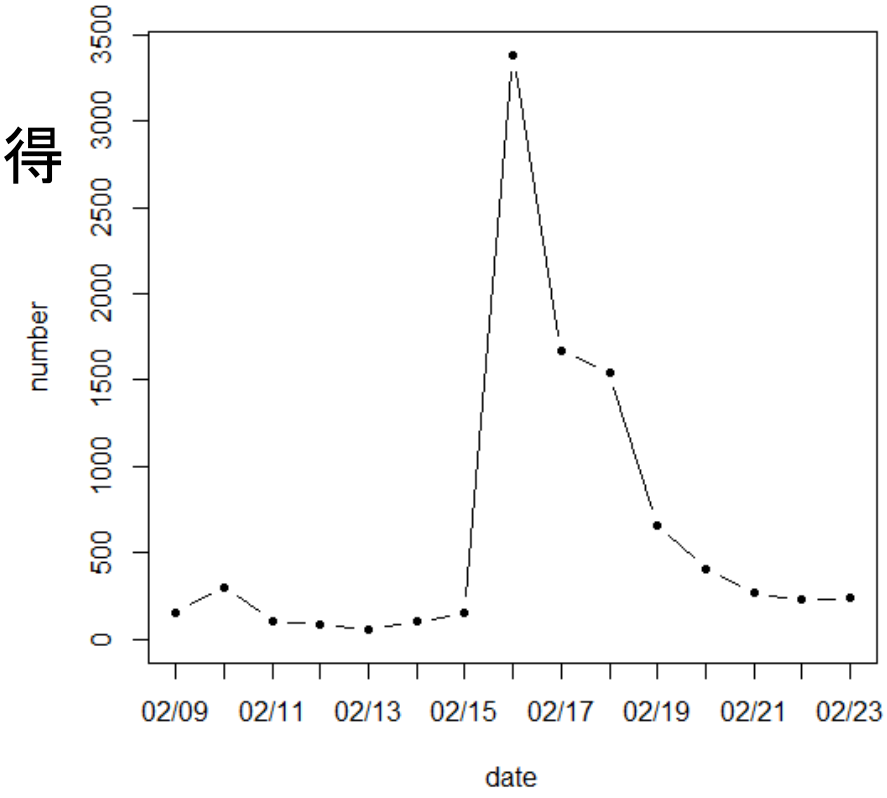
典型的な投稿数のパターン

- イベントの前後で投稿数が対称ではない例

2/16 ソチオリンピックの
ジャンプ男子ラージヒルで
葛西紀明選手が銀メダルを獲得
急激に投稿件数が増加

これらを参考に
数理モデルを提案

「葛西紀明」



＊画像提供：竹村彰通研究室

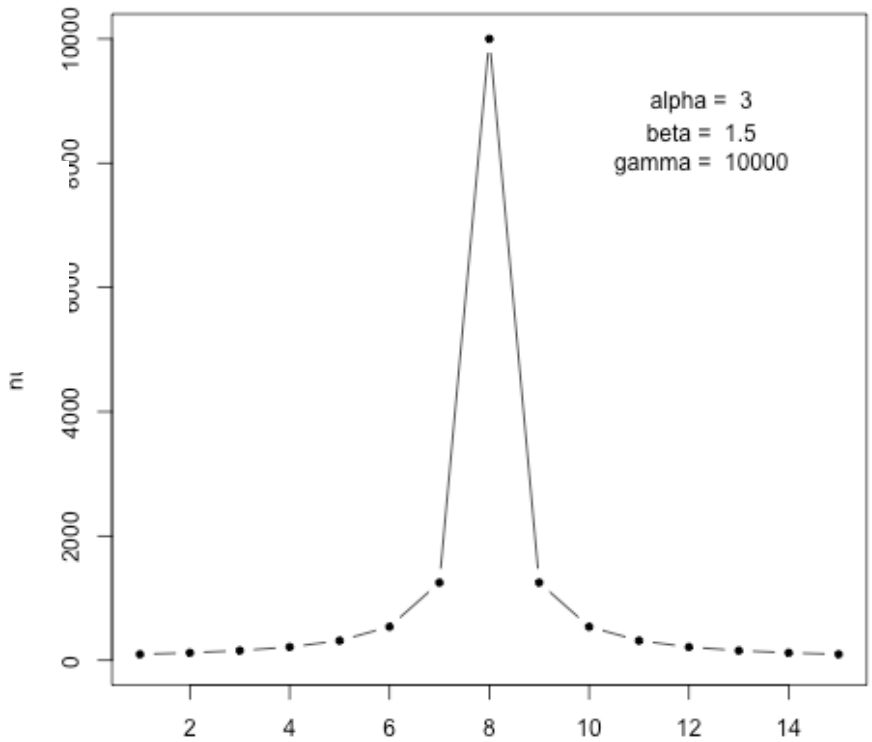
ベキ減衰モデル

- 何らかの関数で投稿数の推移を表したい！

y: 投稿数　t_0:イベント発生日

$$y_t \doteq \gamma \frac{1}{(\alpha|t-t_0|+1)^\beta}$$

α：イベント直前直後の挙動
β：長期的な増加・減衰
γ：イベントのインパクト



＊画像提供：竹村彰通研究室

パラメータの推定法

- βの推定

$\gamma \to$ yの最大値で決め打ち

これより

$$\log(y_t/y_{t_0}) \doteq -\beta \log(\alpha x_t + 1)$$

αが与えられたとき，最小二乗法より，

$$-\beta \doteq \frac{\sum_t \log(y_t/y_{t_0}) \log(\alpha x_t + 1)}{\sum_t (\log(\alpha x_t + 1))^2}$$

パラメータの推定法

➢ αの推定

逆にβが与えられたとき,

$$\left(\frac{y_t}{y_{t_0}}\right)^{-1/\beta} = \left(\frac{y_{t_0}}{y_t}\right)^{1/\beta} \doteq \alpha x_t + 1$$

これより,

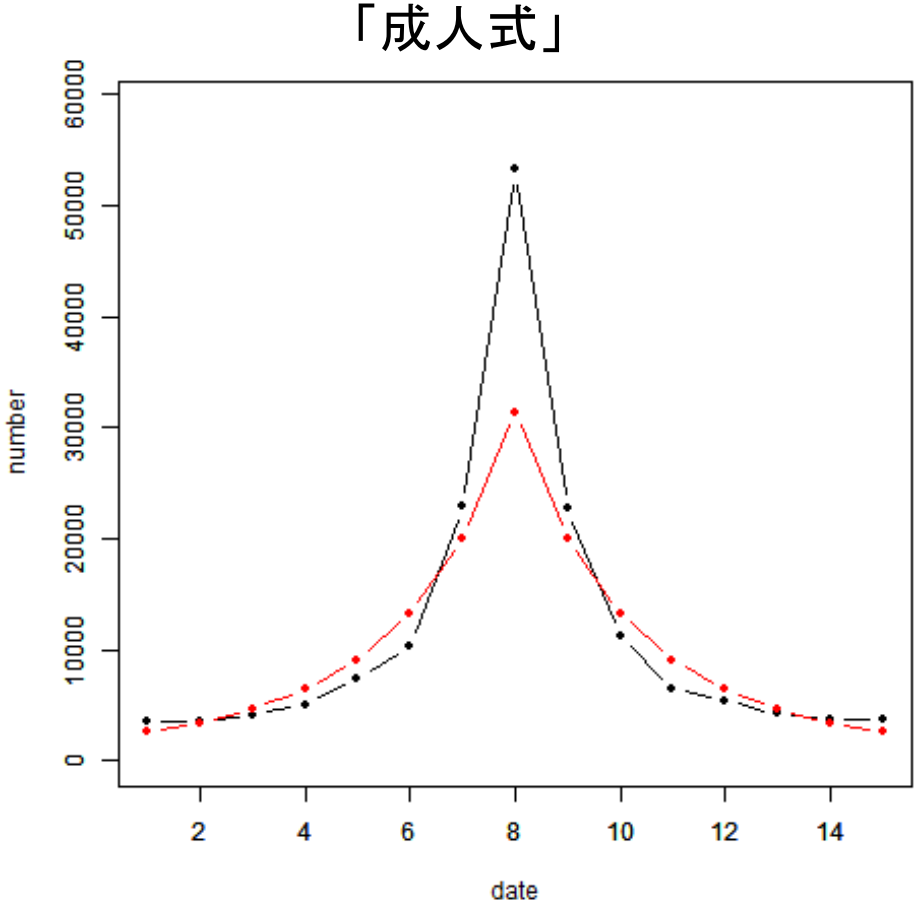
$$\alpha \doteq \frac{\left(\frac{y_{t_0}}{y_t}\right)^{1/\beta} - 1}{x_t}$$

$t = t_0 \pm 1$ とピークの近傍で線形化し,

$$\alpha \doteq \left(\frac{y_{t_0}}{y_{t_0 \pm 1}}\right)^{1/\beta} - 1$$

ベキ減衰モデル

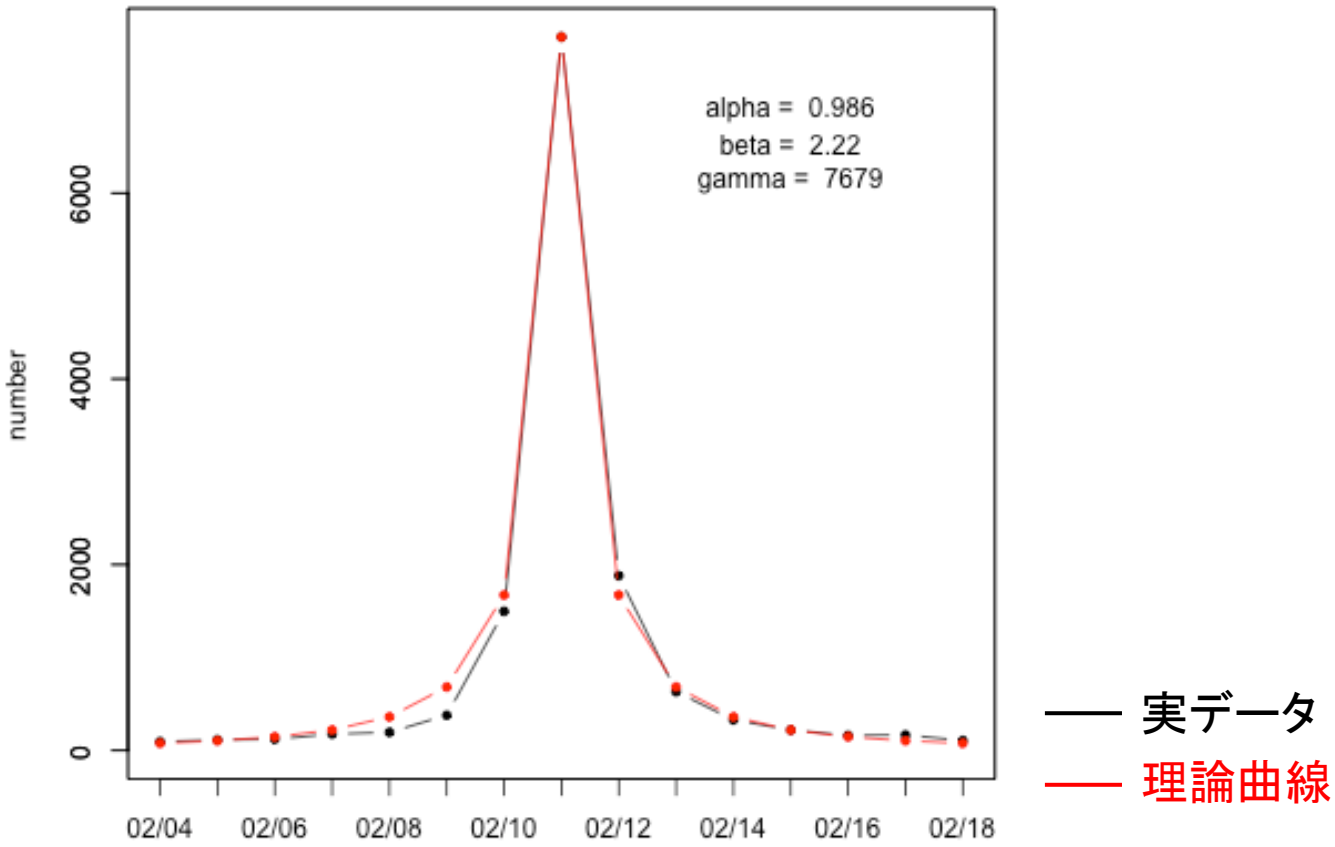
- αの変化に対するグラフの挙動



＊画像提供：竹村彰通研究室

ベキ減衰モデル

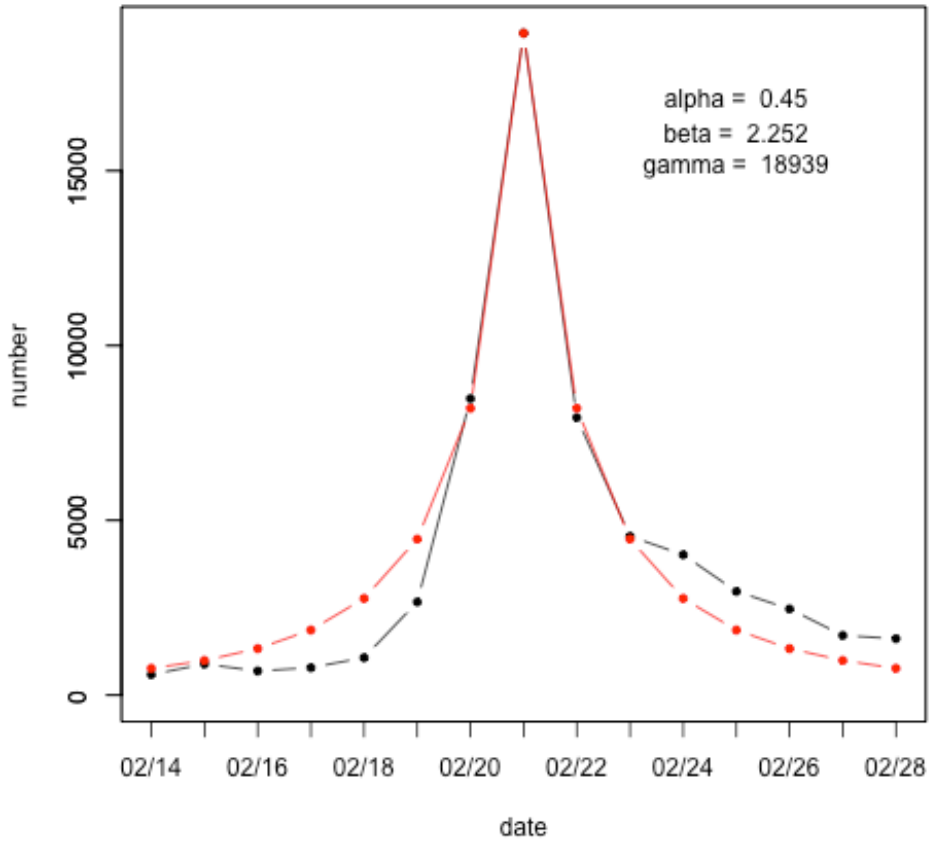
➢ 実際のデータに適用(例:「建国記念日」)



＊画像提供:竹村彰通研究室

ベキ減衰モデル

➢ 実際のデータに適用（例：「浅田真央」）



—— 実データ
—— 理論曲線

＊画像提供：竹村彰通研究室

ベキ減衰モデル

- 諸イベントの特性

	イベント前			イベント後		
キーワード	α	β	γ	α	β	γ
アカデミー賞	25.339	0.463	2923	0.509	0.799	2923
浅田真央	0.171	5.088	18939	1.272	1.059	18939
東日本大震災	8.086	0.883	41566	1.989	1.31	41566
葛西紀明	599709	0.233	3381	0.396	2.121	3381
建国記念日	1.109	2.191	7679	0.852	2.283	7679

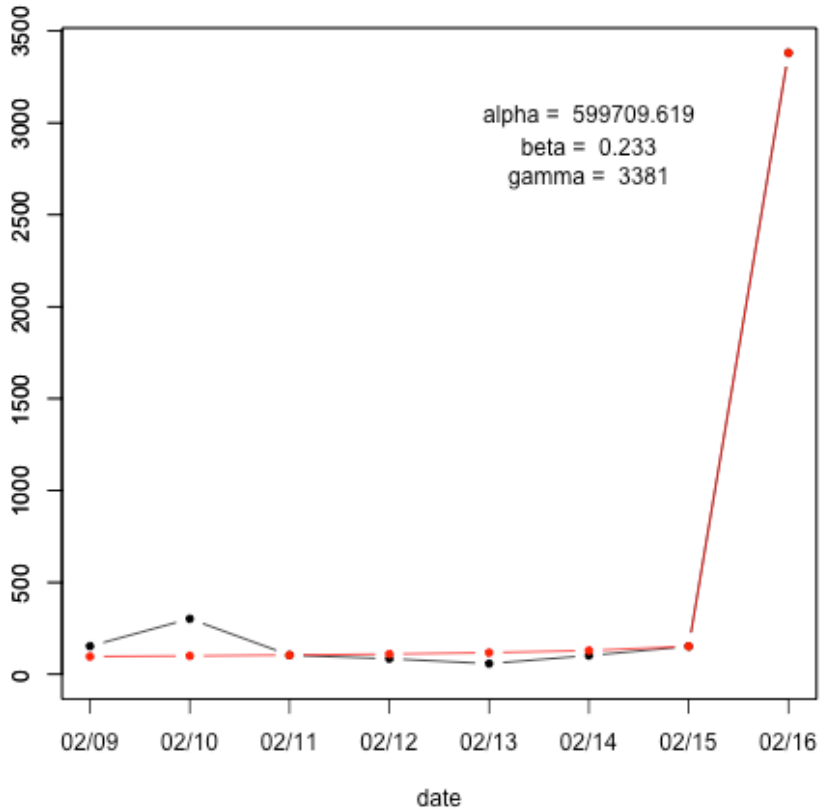
ベキ減衰モデル

- 諸イベントの特性

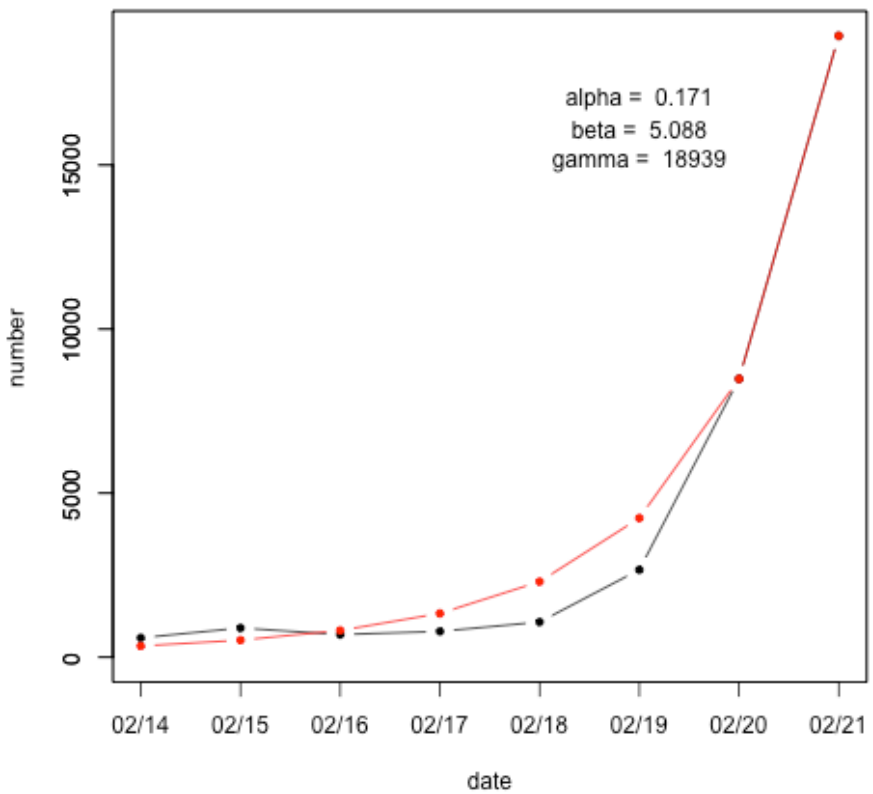
	イベント前			イベント後		
キーワード	α	β	γ	α	β	γ
アカデミー賞	25.339	0.463	2923	0.509	0.799	2923
浅田真央	0.171	5.088	18939	1.272	1.059	18939
東日本大震災	8.086	0.883	41566	1.989	1.31	41566
葛西紀明	599709	0.233	3381	0.396	2.121	3381
建国記念日	1.109	2.191	7679	0.852	2.283	7679

ベキ減衰モデル

➢ 葛西選手 vs 浅田選手



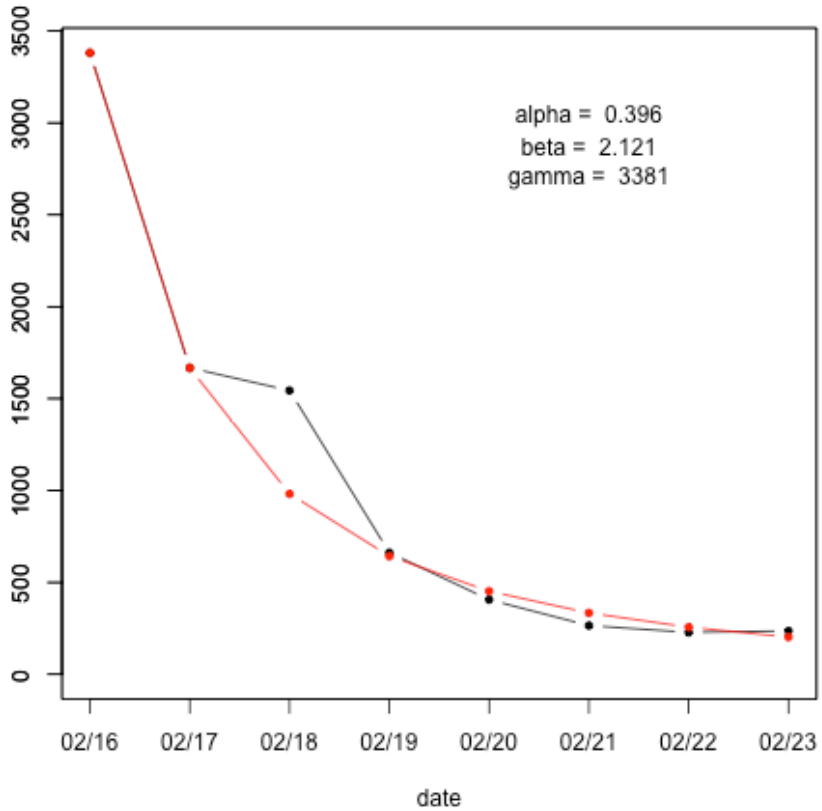
＊画像提供：竹村彰通研究室



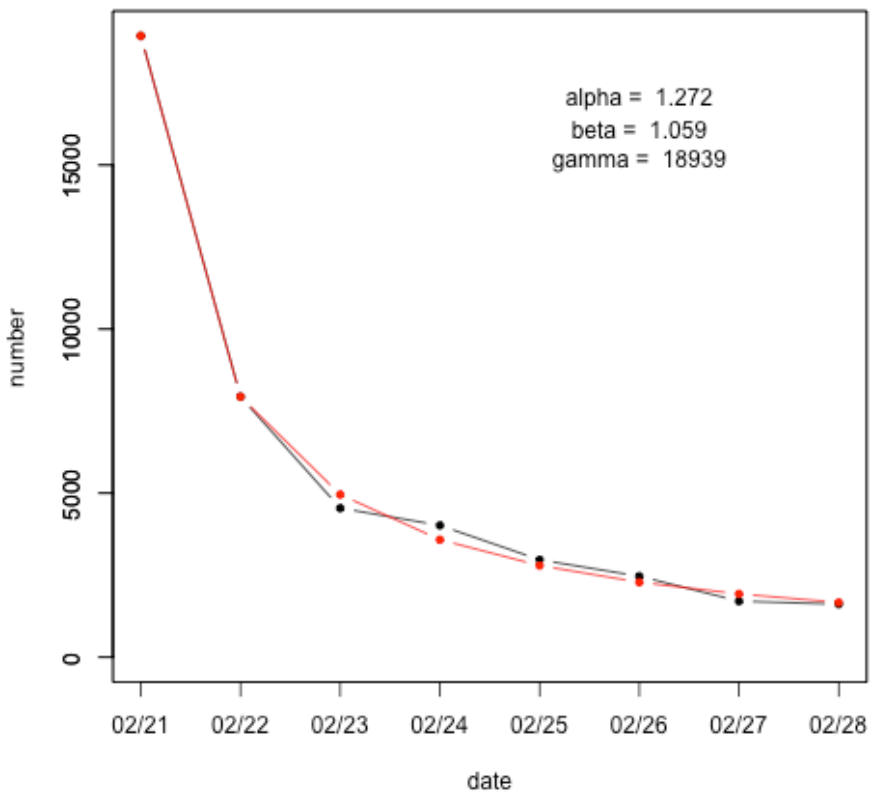
＊画像提供：竹村彰通研究室

ベキ減衰モデル

- 葛西選手 vs 浅田選手



＊画像提供：竹村彰通研究室



＊画像提供：竹村彰通研究室

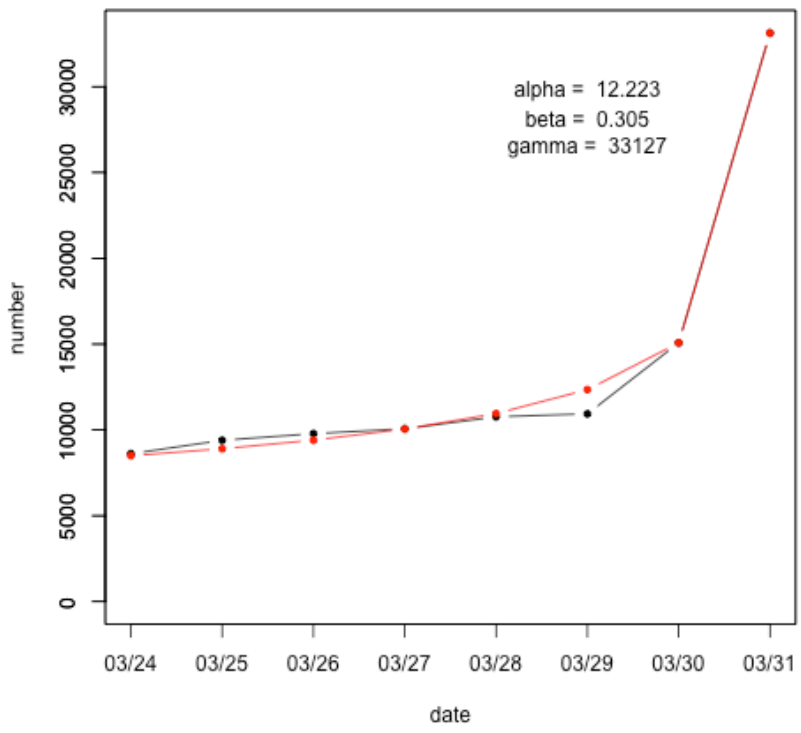
消費増税の話題性の解析

4/1(火)に消費税が8%に増税された.

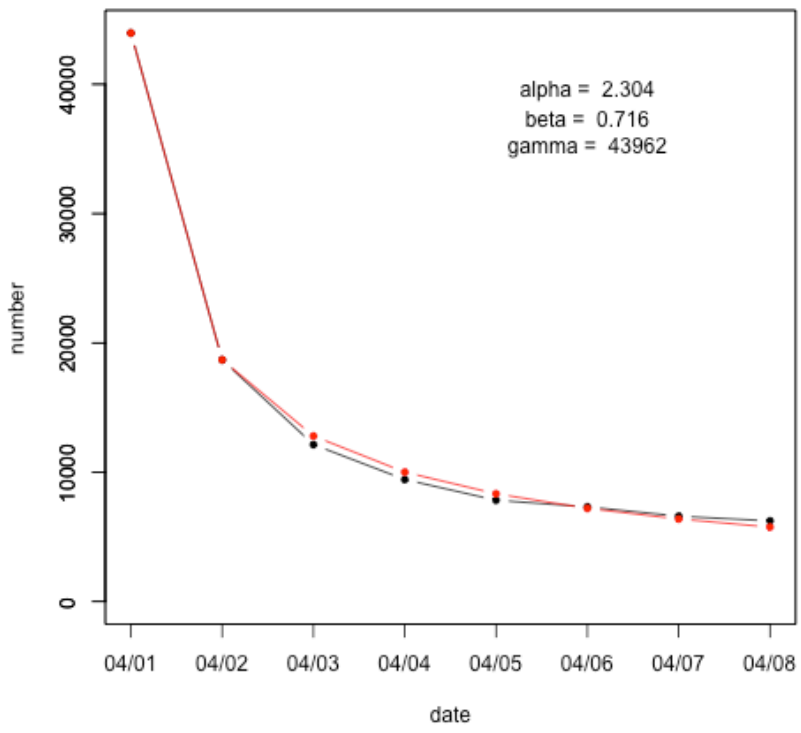
ベキ減衰モデルを用い，消費増税における消費者の購買意識の影響について，
解析を行った.

消費増税の話題性の解析

➢ 「消費税」への関心の推移



＊画像提供：竹村彰通研究室



＊画像提供：竹村彰通研究室

消費増税の話題性の解析

- 男女差

	増税前			増税後		
	α	β	γ	α	β	γ
全体	12.223	0.305	33127	2.304	0.716	43962
男性	6.171	0.426	5105	2.267	0.774	7255
女性	13.262	0.364	6530	1.842	0.934	8397

消費増税の話題性の解析

- 男女差

	増税前			増税後		
	α	β	γ	α	β	γ
全体	12.223	0.305	33127	2.304	0.716	43962
男性	6.171	0.426	5105	2.267	0.774	7255
女性	13.262	0.364	6530	1.842	0.934	8397

消費増税の話題性の解析

- 地域差

	増税前			増税後		
	α	β	γ	α	β	γ
北海道	3.287	0.544	351	3.561	0.698	479
東北	9.861	0.416	378	1.702	0.991	506
関東	30.141	0.283	2833	2.02	0.853	3902
中部	7.901	0.418	1059	2.083	0.876	1402
近畿	19.535	0.316	1306	2.487	0.813	1810
中国	6.769	0.44	340	3.403	0.752	442
四国	10.405	0.357	174	1.801	0.952	256
九州・沖縄	76.174	0.266	541	2.127	0.836	700

消費増税の話題性の解析

- 地域差

	増税前			増税後		
	α	β	γ	α	β	γ
北海道	3.287	0.544	351	3.561	0.698	479
東北	9.861	0.416	378	1.702	0.991	506
関東	30.141	0.283	2833	2.02	0.853	3902
中部	7.901	0.418	1059	2.083	0.876	1402
近畿	19.535	0.316	1306	2.487	0.813	1810
中国	6.769	0.44	340	3.403	0.752	442
四国	10.405	0.357	174	1.801	0.952	256
九州・沖縄	76.174	0.266	541	2.127	0.836	700

消費増税の話題性の解析

- 消費財ごとの違い

	増税前			増税後		
消費材	α	β	γ	α	β	γ
トイレットペーパー	8.551	0.441	946	1.468	1.155	656
タバコ	33.32	0.316	455	2.792	0.752	594
定期券	2505.6	0.209	410	1.338	1.479	474

消費増税の話題性の解析

- 消費財ごとの違い

	増税前			増税後		
消費材	α	β	γ	α	β	γ
トイレットペーパー	8.551	0.441	946	1.468	1.155	656
タバコ	33.32	0.316	455	2.792	0.752	594
定期券	2505.6	0.209	410	1.338	1.479	474

消費増税の話題性の解析

- 消費財ごとの違い

	増税前			増税後		
消費材	α	β	γ	α	β	γ
トイレットペーパー	8.551	0.441	946	1.468	1.155	656
タバコ	33.32	0.316	455	2.792	0.752	594
定期券	2505.6	0.209	410	1.338	1.479	474

消費増税の話題性の解析

- 消費財ごとの男女差

	増税前			増税後		
ガソリン	α	β	γ	α	β	γ
全体	5.02	0.728	2932	2.739	0.945	3064
男性	2.947	0.88	636	2.523	0.931	656
女性	6.792	0.842	597	3.52	1.042	597

消費増税の話題性の解析

- 消費財ごとの男女差

	増税前			増税後		
ガソリン	α	β	γ	α	β	γ
全体	5.02	0.728	2932	2.739	0.945	3064
男性	2.947	0.88	636	2.523	0.931	656
女性	6.792	0.842	597	3.52	1.042	597

まとめと今後の課題

成果

- ブログの投稿数の推移を近似する，ベキ減衰モデルの提案
- 関心の高まり・薄まりを表す指標の作成

今後の課題

- 何らかの原理に基づく，ベキ減衰モデルの導出
- 確率モデルの構築

実社会における応用

- Web広告の効果改善
- 消費税10%化への準備

このモデルを統計モデルに発展させたい

- まだ「統計モデル」になっていない.

統計モデル＝固定効果＋確率的な誤差

において，固定効果のところしかモデル化していない.

- そのため，推定値の信頼区間が作れない.

- 今後統計モデルに発展させる予定.

3. ビッグデータ時代の統計学

(最近いろいろと書いてはいるのですが，実はなかなか考えがまとまりません)

データ量の増加　(wikipedia等から)

- メガバイト (10^6): 3.5 インチフロッピー
- ギガバイト (10^9): cd-rom, dvd
- テラバイト (10^{12}): 現在の普通のハードディスク
- ペタバイト (10^{15}): アメリカにある図書館全体の情報
- Exabyte (10^{18}): 人類がこれまで発した言葉の全体
- Zettabyte (10^{21}): 2012年中のインターネットトラフィック

ビッグデータ分析のニーズ

大量のデータを分析するのか？

- ビジネス: 利益に通じるなら
- 製造業: 品質の向上や新しい製品開発につながるなら
- 科学: 新しい知見が得られるなら

誰が分析するのか？

- 「データサイエンティスト」の世界的不足
- コンピュータとアルゴリズムがある程度自動的に分析「機械学習」

→ 新たな分析手法の開発

→ でも解釈するのは人間

結果の解釈，信頼性の評価

- データマイニングの「成功例」

*

*

新たな分析手法の開発

- テキストデータなどの非構造化データの解析手法
- 画像データ，動画データ，の増大

- 数理モデルや統計モデルの開発
- モデルの推定法
- モデル選択
- 最適化等のアルゴリズムの改良，実装の効率化
- オンラインアルゴリズム (データが来るごとに処理，更新)

- プライバシーの問題

データと情報

- データとしては大量でも，情報はほとんどないものもある.
- 例: サイコロを投げ続けて出た目を記録して行く: 1,5,2,4,3,…
- ハードディスク一杯になるまで記録したら「ビッグデータ」
- このデータは有用？
  - 常識的な答: 単にデタラメを記録しただけだから無意味.
  - 「サイコロを投げて出た目でディスクをうめただけ」と言えば済む．情報とは言えない.
  - 「データの生成過程」の記述のほうが重要.

    → 統計学の考え方
  - 1がたまたま27回続けて出たら珍しいので記録の価値があるか？

データと情報

- デタラメなデータほど情報量が多いとする考え方もある.

  → コルモゴロフ複雑度

- でたらめなデータ (1,5,2,4,3,...) はそのまま記録するよりない. 簡単に記述できない.
- パスワードはでたらめでないと困る(12345678とかは駄目)
- たんさくのパスワードは覚えられないので, 皆困っている.
- 1,1,1,1,.... とハードディスクを1で埋め尽くすのはもちろん無意味.
- どちらにせよ, 多量のデータ自体が情報を含むわけではない.
- でも何が有用な情報かはわからない.

監視カメラや衛星が記録する世界

- 監視カメラが人々の出入りを記録している.
- 人々の移動はランダム.
- ランダムな現象をそのまま記録しても意味があるのか.
- 通常の監視カメラは，２週間程度で画像を消去.
- しかし犯罪がおきるととたんに有用となる.
- 画像解析技術が進歩して，個人が自動的に識別されるようになったとして，プライバシーの問題はどうなるか

- 天気予報も解像度があがり，局所的な予報が可能になってきた.
- 短期的な予報の精度もあがってきた
- それでも，いくらデータあっても，今年の夏が暑いかどうかは予測できない
- 数理モデルは重要，データは必ずしも重要でない (しつこい)

SNSのデータから何がわかるのか

- ブログや twitter のデータはほとんどがノイズ
  「おなかがすいた」「今日の講義つまらない」等
- 発言している人々の属性が把握できない.
- 発言数の分布の裾が重い.
- 動向をとらえることはできるが，確定的なことを言うのは難しい
- 差分には情報があるが，水準には意味がない

- 「母集団」が設定できない.
- 統計的手法の根拠: 母集団，あるいはより仮想的な統計モデルの安定性

統計モデル＝固定効果＋確率的な誤差

ビッグデータ時代における
個人の情報と集団の情報

- ビッグデータ時代は「個人情報」の時代.
- ネット経由で買い物をすると履歴が残り，また他のユーザを参考にして商品が推薦される.
- 利便性のために，消費者が個人情報を提供している.
- 一方で，国のおこなう統計調査への非協力が増えており，統計調査の環境が悪化している.
- 国勢調査などの国の基本的な調査で調べる情報は，利便性と引き換えに得られる情報ではない.
- 統計: 集団に関する情報
    - 個人の情報 ↔ 小集団の情報 ↔ 集団の情報

データサイエンティストの育成と統計リテラシーの向上

- ビッグデータの時代にデータ解析に専門性を有する人材が重要.
- 「サイエンス」には違和感あるが，ネーミング.
- 数学もアルゴリズムも統計もビジネスもできる人材.
- 専門性を有する人材のみならず，一般的な統計的なリテラシーの向上も必要