新型コロナ関連用語の受容度に関するシンプルな統計モデリング
Simple statistical modelling of the acceptance of COVID-19 related terms

横山詔一[1]、相澤正夫[1]、田中牧郎[2]，久野雅樹[3]、田中祐輔[4]

1国立国語研究所、2明治大学、3電気通信大学、4青山学院大学

20230404 NINJALサロン 15時10分～16時10分

サロンの討論を参考に「スティホーム」→「ステイホーム」、「残滓の蓄積」に取り消し線20230406

1. きょうの原データについて
2. なぜいま、この研究か（学術的動機）
3. なぜいま、この研究か（社会的動機）
4. どのようなデータか
5. 「この言葉をそのまま使うのがいい」と回答した割合を、その語に対する「受容度」だと定義する
6. 受容度のグラフを目視して頭に浮かんだこと
7. グラフを目視したら、次は「現時点での科学の掟」に従って結果を分析する
8. どうする分析
9. GLMMは「生態学的誤謬」に対応可能
10. まとめと今後の課題
11. 参考：社会を対象としたデータについて
12. 引用・参考文献
13. 【付録】国立国語研究所の研究公正研修「ケーススタディ」（2023年3月9日）で話題になったQRP（Questionable Research Practice：疑わしい研究行為）の防止に向けて：いわゆる「*p*値ハッキング」に関する自己チェックリストの一部分を試作してみました

きょうの原データについて

きょうは文化庁の「国語に関する世論調査」を取り上げます

https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/kokugo_yoronchosa/index.html

以下の『令和2年度 国語に関する世論調査 〔令和3年3月調査〕 』報告書の調査集計表（オープンデータ）を分析します

https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/kokugo_yoronchosa/pdf/93710501_01.pdf

なぜいま、この研究か（学術的な動機）

言語意識研究の絶好の機会

- 新型コロナウイルス感染症に関連した用語については、性別や年代（幼児は除く） に関係なく、ほぼすべての人が一気に高頻度で接触したと考えられる
- そのため、純接頻度の男女差や年代差は、前例のないくらい極めて小さいものだったと仮定できそうである
- ここでは、このような特別な条件下で収集された全国規模の大量のデータに基づいて、新型コロナウイルス感染症関連用語の「受容度（acceptance） 」を調べる

なぜいま、この研究か（学術的な動機）

言語意識研究の絶好の機会

- 分析対象とする語は、単なる新語・流行語の類ではなく、生存に関わる重要語として短期間で世の中に流通したと想定されるものである
- 本研究は、回答者の属性による受容実態の違いを明らかにする。あわせて、その結果の背後に潜在するメカニズムを簡便な手法で捉え、受容確率の予測を試みる
- 受容（acceptance）は定着（establishment）の前提あるいは必要条件だと考える。たとえば、「仏教文化は古代の日本社会に受容され、やがて定着していった」と言える

- 受容度と定着度の違いについては以下が参考になる

久屋愛実（2021）「「国語に関する世論調査」に見る外来語の動態―外来語を考える四つの視点―」『日本語学』40(2):84-94, 明治書院

なぜいま、この研究か（学術的な動機）

参考文献

林廷修（2021）「外来語とその言い換え表現からみた言葉の受け入れに関する研究―新型コロナウイルス関連用語を例に―」『日本語学会2021年度春季大会予稿集』61-66

- 2020年1月～12月の毎日新聞本文からコロナ関連の外来語を抽出し、「マスク、ウイルス、ワクチン」のような広く普及しているものを除外して抽出語を次のように4分類した
- 外来語のみ：例「オーバーシュート」【L1】、外来語（言い換え表現）の形：例「オーバーシュート（感染爆発）」【L2】、言い換え表現（外来語）の形：例「感染爆発（オーバーシュート）」【N1】、言い換え表現のみ：例「感染爆発」【N2】
- 次に、L1とL2を外来語、N1とN2を言い換え表現と見なし、外来語と言い換え表現を合わせた出現数が30例以上のもの8語を抜き出した
- その8語に関する日本語母語話者のイメージを探るため、形容詞ペア8つ（例「軽いー重い」「馴染みがあるー馴染みがない」「分かりやすいー分かりにくい」）に対する評定データを収集した（調査対象者数62名、2021年1月15日～31日、オンライン調査）

なぜいま、この研究か（社会的な動機）

公共の言語問題に対する研究の系譜：国立国語研究所（特に独法後）の成果を中心に

2001年から2009年まで

- 「外来語」言い換え提案：https://www2.ninjal.ac.jp/gairaigo/
- 「病院の言葉」を分かりやすくする提案：https://www2.ninjal.ac.jp/byoin/

2010年から2020年までの間をどうみるか、休眠？→ 社会との接点の（積極的な）回復が急務：学術と社会の関係が厳しく見直しを迫られている

2021年から

- 難解な感染症関連用語の言い換えや説明の案出と理解促進効果の検証（横山科研）
- 多言語・多文化社会における言語問題に関する研究（PJリーダー：朝日祥之）

どのようなデータか

代表性を有するデータ

- 全国規模
- 16歳以上の個人
- ランダムサンプリング：層化2段階抽出法
- 調査対象者数6,000名、有効回収数3,794名（回収率63.2％）

方法

- ネット調査ではなく、郵送法（前年度までは面接聴取法）
- 調査実施機関は中央調査社（1954年（昭和29年）設立、時事通信社調査室と（旧）国立世論調査所が母体。日本人の読み書き能力1948年調査とも人材面で関連が深い）
  https://www.crs.or.jp/backno/No671/6711.htm
- 調査時期は2021年3月4日から同年3月29日まで → 新型コロナ国内感染の報道開始からほぼ1年経過の時期

報告書の22ページから26ページまで

質問と選択肢

問４　ここに挙げた（１）～（８）の言葉の使われ方について、どのように思いますか。あなたのお考えに最も近いものをそれぞれ一つずつ選んでください。

（１）コロナ禍、（２）ソーシャルディスタンス、（３）３密、（４）濃厚接触、（５）クラスター、（６）不要不急、（７）ステイホーム、（８）ウイズコロナ

- この言葉をそのまま使うのがいい
- この言葉を使うなら、説明を付けたほうがいい
- この言葉は使わないで、ほかの言い方をしたほうがいい

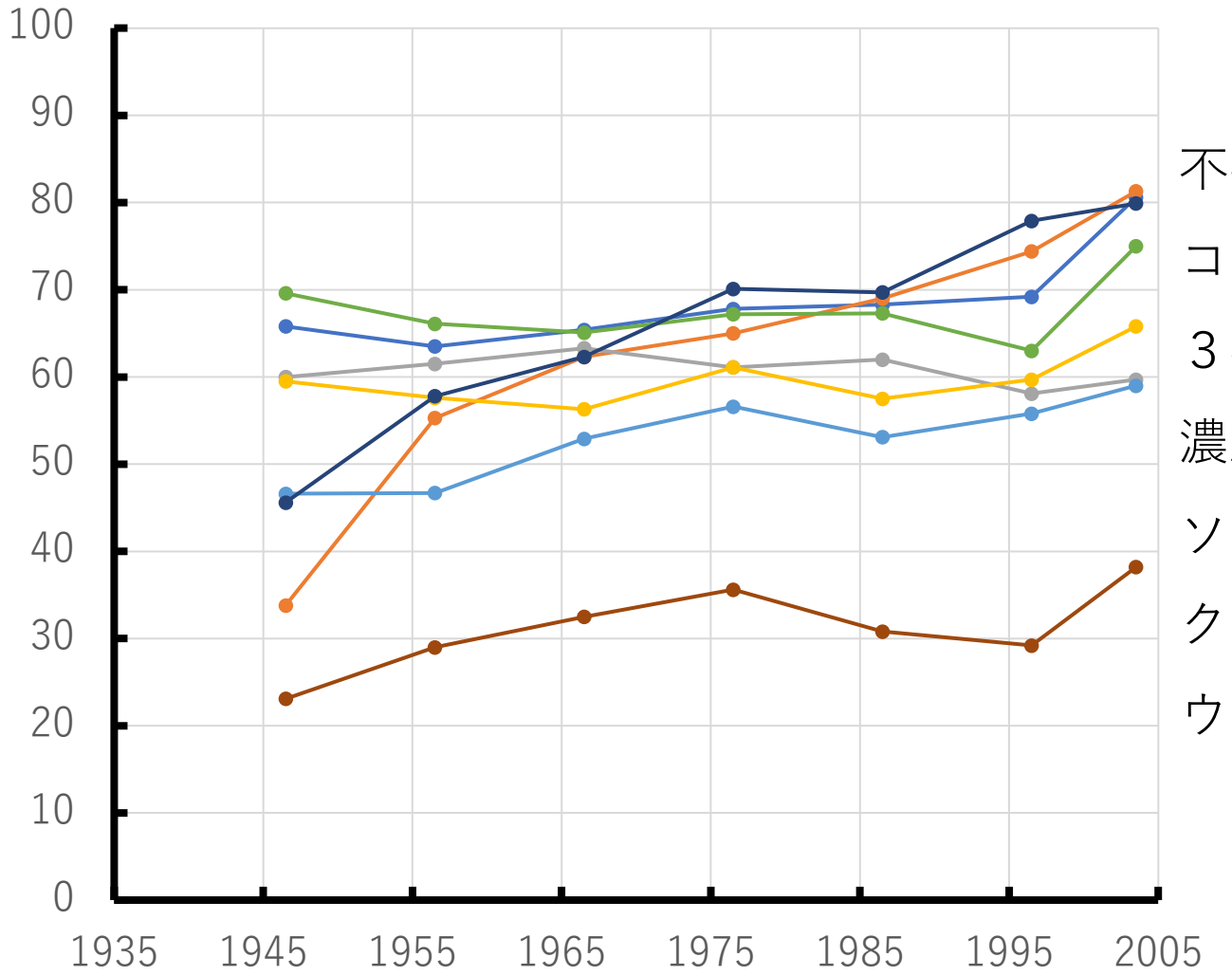
注）選択肢の2番目は「言い添え派」つまり言葉の補助輪が必要だという意見で、３番目は「言い換え派」

「この言葉をそのまま使うのがいい」と回答した割合を、その語に対する「受容度」だと定義する

まず、報告書26ページの集計表からグラフを作成して目視する（地域差については割愛）

- 受容度（縦軸）と生年（横軸）のグラフを作成
- 集計表に生年は示されていないので、調査年から年齢を引き算して求める
- 本研究で、年齢ではなく生年を横軸にする理由は、この先の経年調査、すなわち実時間調査を念頭に置いているから
- 以下、生年を「年代」の意味で使うことがある。生年が早いほうが高年層、遅いほうが若年層となる。グラフでは横軸の左側が高年層、右側が若年層である

「この言葉をそのまま使うのがいい」と回答した割合を、その語に対する「受容度」だと定義する



受容度のランキング

不要不急（67.2%）

コロナ禍（66.8%）

３密＝ステイホーム（61.1%）

濃厚接触（58.9%）

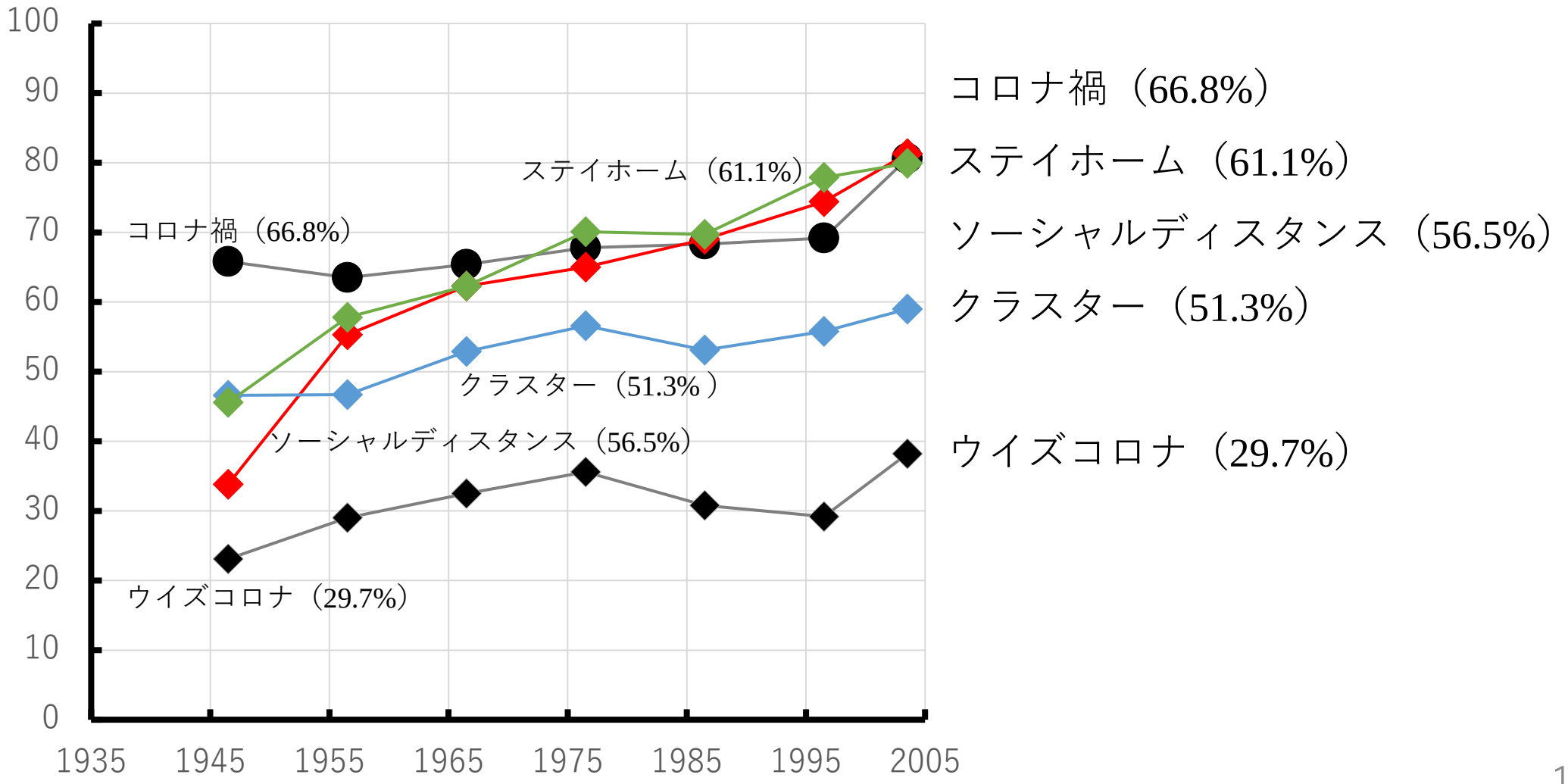
ソーシャルディスタンス（56.5%）

クラスター（51.3%）

ウイズコロナ（29.7%）

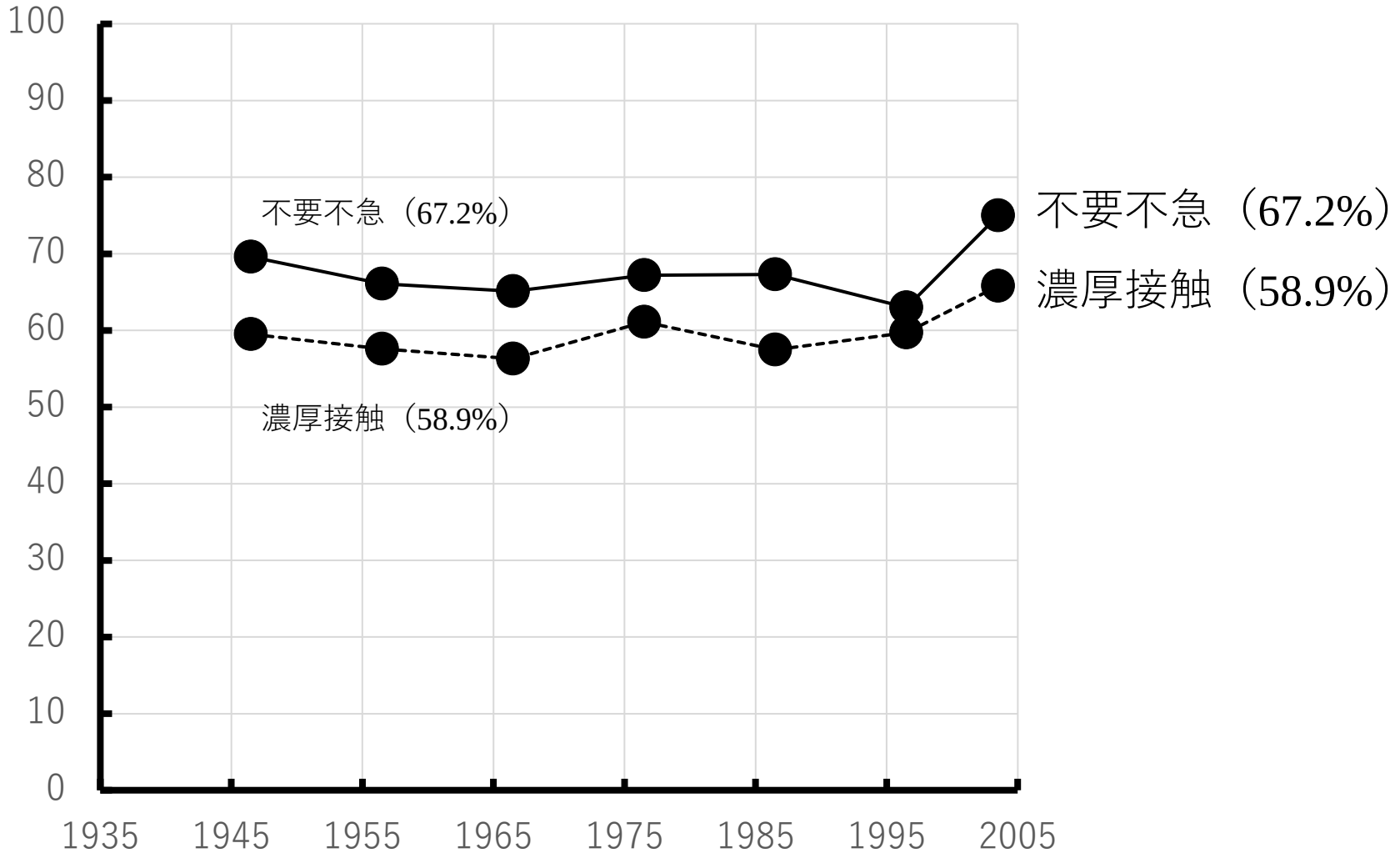
カタカナ表記の4語と「コロナ禍」

- 「ソーシャルディスタンス」は年代差が大きい、「ウイズコロナ」は受容度が最低



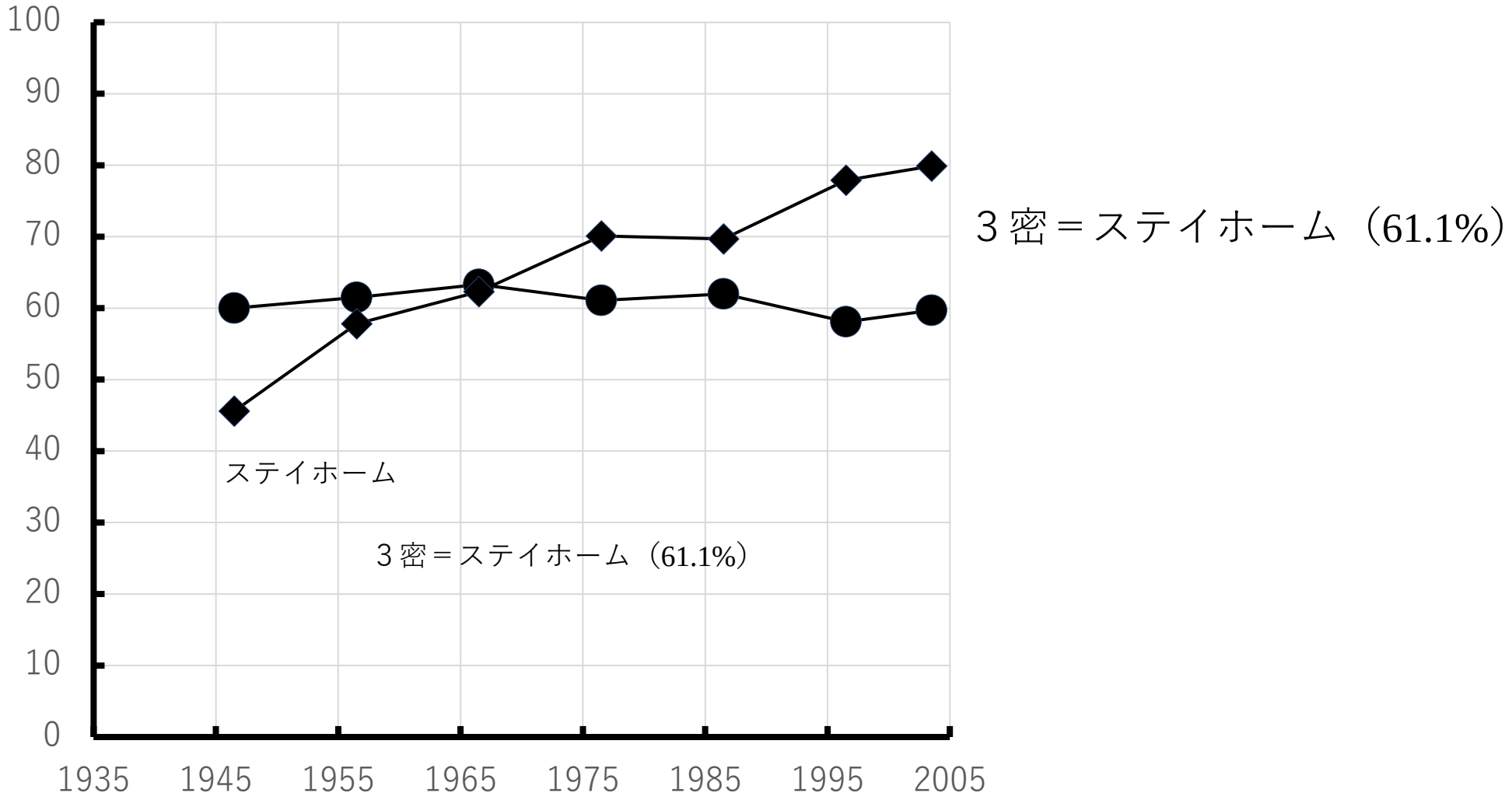
「不要不急」と「濃厚接触」は年代差がないような、あるような・・・

- 目視だけだと判断に迷うケース



「３蜜」のグラフは平坦、「ステイホーム」は年代差あり

- 全体で見ると受容度は同じ



受容度のグラフを目視して頭に浮かんだこと

- 文化庁国語課の2021年度報告書には年代差、性差、地域差について要点がまとめられている。統計的検定などは登場しないが、記述の内容は妥当であり、説得力を有する
- しかし、報告書に述べてあることはデータから得られる情報の一部だということも容易に想像できる
- 学術的な知見や手法を動員してデータから有益な情報をなるべく多く搾り出し、論文にまとめて公表するのがよい。その代表例が以下である

久屋愛実 （2021）「英語由来語彙を公共コミュニケーションでどう運用するべきか：全国調査のデータを活用した福祉言語学的考察」『計量国語学』33巻3号, 130-145

https://www.jstage.jst.go.jp/article/mathling/33/3/33_130/_article/-char/ja/

受容度のグラフを目視して頭に浮かんだこと

論文執筆の際にポイントとなりそうなアイデアその１

- 新語が徐々に世間に広がることにより年代差が生じると考えられている。しかし、今回はどの年代も同時に一気に同じ語に接触しているにもかかわらず、カタカナ語で年代差が生じるのはなぜか？→本当に差があるのか、統計的分析で確認してみる（以下同様）
- 「ソーシャルディスタンス」＝ソーシャル＋ディスタンスで、高年齢層はその要素の語に不慣れなため年代差が大きいのではないか（グラフの傾きが大きい）？
- ほかに比べて「ウィズコロナ」は世に出るのが遅かったので受容度が低く、年代差も小さい（グラフの傾きが小さい）のではないか？
- 同一年代でも、日常生活で外来語が接触語彙や理解語彙である人と、そうでない人では受容度に差があるのでは？

受容度のグラフを目視して頭に浮かんだこと

論文執筆の際にポイントとなりそうなアイデアその２

- 「不要不急」はコロナ禍より前からマスメディア等に登場していた（台風接近時など）ので新語とは言えないのでは？
- 「濃厚接触」はいつごろからマスメディアに登場したのだろうか？

受容度のグラフを目視して頭に浮かんだこと

論文執筆の際にポイントとなりそうなアイデアその３

- 久屋（2016）が外来語の研究で男女差が明確に出ることを指摘しているが、今回も同じような結果になるのだろうか？ → 本当に男女差があるのか、統計的分析で確認してみる

久屋愛実（2016）「見かけ上の時間を利用した外来語使用意識の通時変化予測」『日本語の研究』12巻4号,69-85

https://www.jstage.jst.go.jp/article/nihongonokenkyu/12/4/12_69/_article/-char/ja/

受容度のグラフを目視して頭に浮かんだこと

論文執筆の際にポイントとなりそうなアイデアその４

- 数表をざっと見ると、全体的に地域差は小さいように見えるが、ごく一部の語、例えば「濃厚接触」では四国と東北で受容度にかなりの差（16ポイント強）があり、ひときわ目立っている。その理由は？
- 回収率に5%ほどの男女差がある。この影響をどう考えるか？

グラフを目視したら、次は「現時点での科学の掟」に従って結果を分析する

今回の8語をタイプ分けする。目視だけに頼るのではなく、統計モデリングを活用

統計モデリングとは何か？

- 統計モデリングという用語は、なにか特定の手法を指しているのではない。たとえば、久保（2019）の通称緑本では「モデルを作って観測データにあてはめて現象を理解する」方法が統計モデリングだと言う
- これだけの説明だと何だかよく分からない
- そこで、本研究では久保（2019）の説明を次のように改変してみた
- 本研究における統計モデリングとは「ある事象の生起確率を予測するロジスティック回帰モデルを作って観測データにあてはめて現象を理解する」こと

グラフを目視したら、次は「現時点での科学の掟」に従って結果を分析する

今回の8語をタイプ分けする。目視だけに頼るのではなく、統計モデリングを活用

- ただし、必要最低限のシンプルな統計モデリングをおこなう
- 今回は、性と生年（年代）の2変数で受容度を予測するロジスティック回帰モデルを構築する。当然のことながら変数の有意差検定もおこなう
- 男女差（あり／なし）と年代差（あり／なし）の組み合わせで4つのタイプができるので、それを用いて語を分類してみる
- 具体的には8語を以下の4タイプのいずれかに分類する。男女差があって年代差もある（タイプAと称する）、男女差だけある（タイプB）、年代差だけある（タイプC）、男女差も年代差もない（タイプD）
- GLMM（一般化線形混合モデル）も試してみる

どうする分析

GLM（一般化線形モデル）のロジスティック回帰分析

- 性と生年（年代）の2変数から受容度を予測するモデルを語別に作成

	生年		性		定数	AIC
	係数	有意水準	係数	有意水準		
コロナ禍	0.005932	<1%	0.230844	<0.1%	-11.092796	100.13
ソーシャルディスタンス	0.034403	<0.1%	0.292903	<0.1%	-67.542429	134.93
３密	-0.0003872	> 5%	0.1986834	<1%	1.1057934	98.524
濃厚接触	0.001419	> 5%	-0.041566	> 5%	-2.407982	96.443
クラスター	0.008689	<0.1%	0.324404	<0.1%	-17.215852	99.65
不要不急	-0.001442	> 5%	-0.103696	> 5%	3.611447	103.59
ステイホーム	0.027383	<0.1%	0.386338	<0.1%	-53.589912	105.44
ウイズコロナ	0.008902	<0.1%	0.335409	<0.1%	-18.564394	114.99

どうする分析

GLM（一般化線形モデル）のロジスティック回帰分析

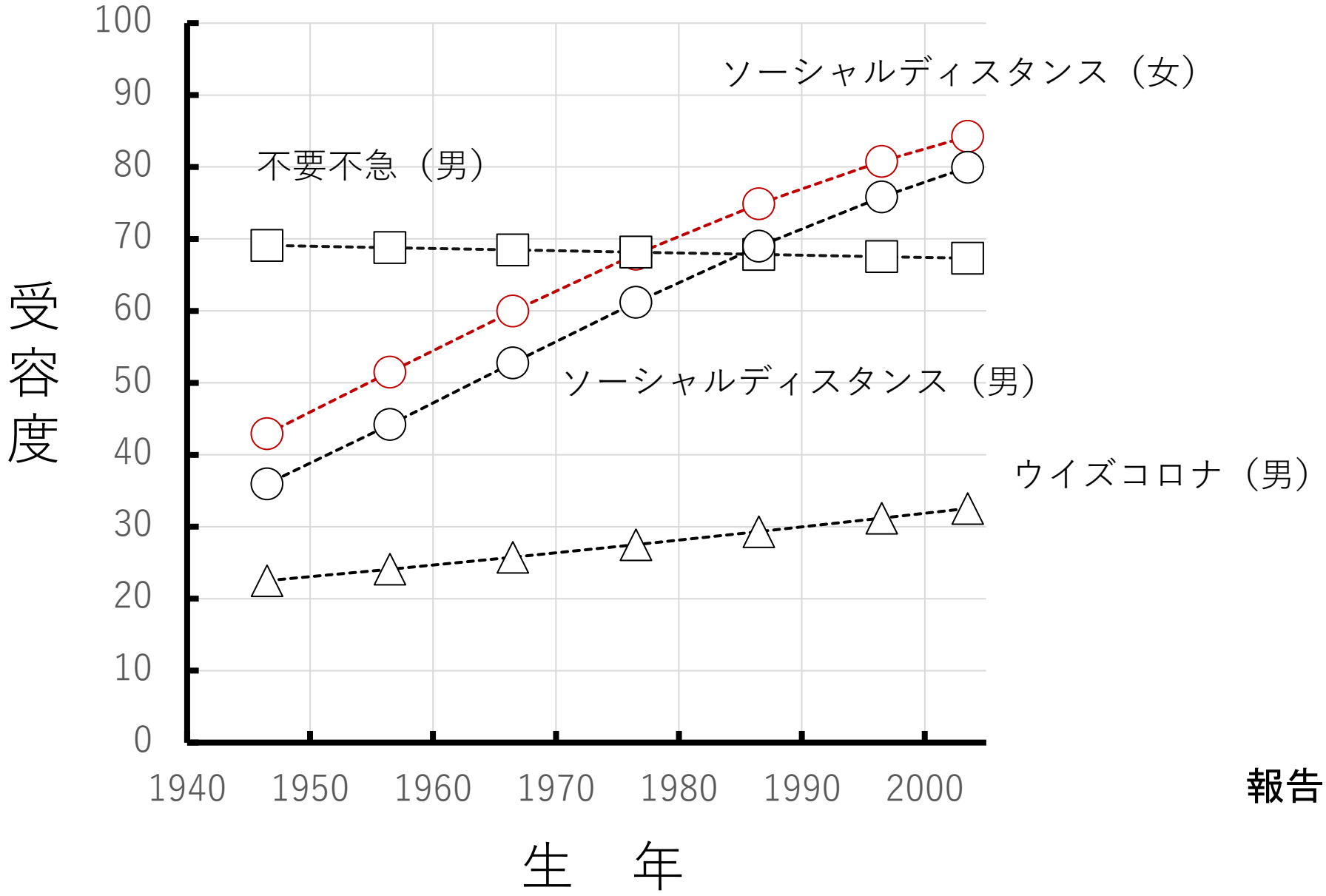
カタカナ表記語はタイプA

- 先に示したタイプA（男女差と年代差の両者が統計的に有意）：ソーシャルディスタンス、クラスター、ステイホーム、ウイズコロナの4語（カタカナ）とコロナ禍の1語（カタカナ＋漢字）。女性＞男性、若年層＞高年層
- タイプB（男女差のみ統計的に有意）：３密の1語（アラビア数字や漢数字＋漢字）。女性＞男性
- タイプC（年代差のみ統計的に有意）：なし
- タイプD（男女差も年代差もなし）：濃厚接触、不要不急の2語（漢字）

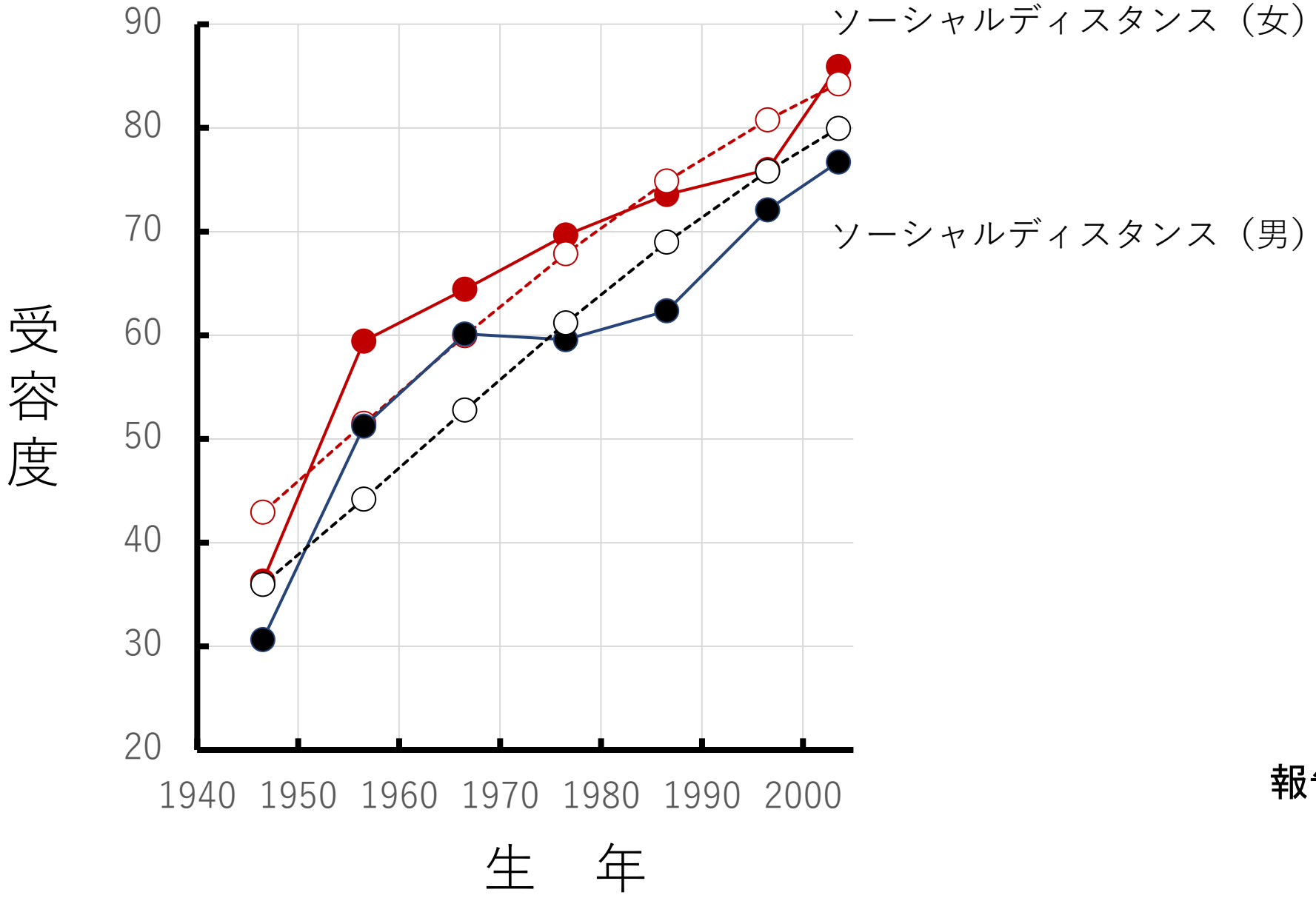
どうする分析

GLM（一般化線形モデル）のロジスティック回帰分析

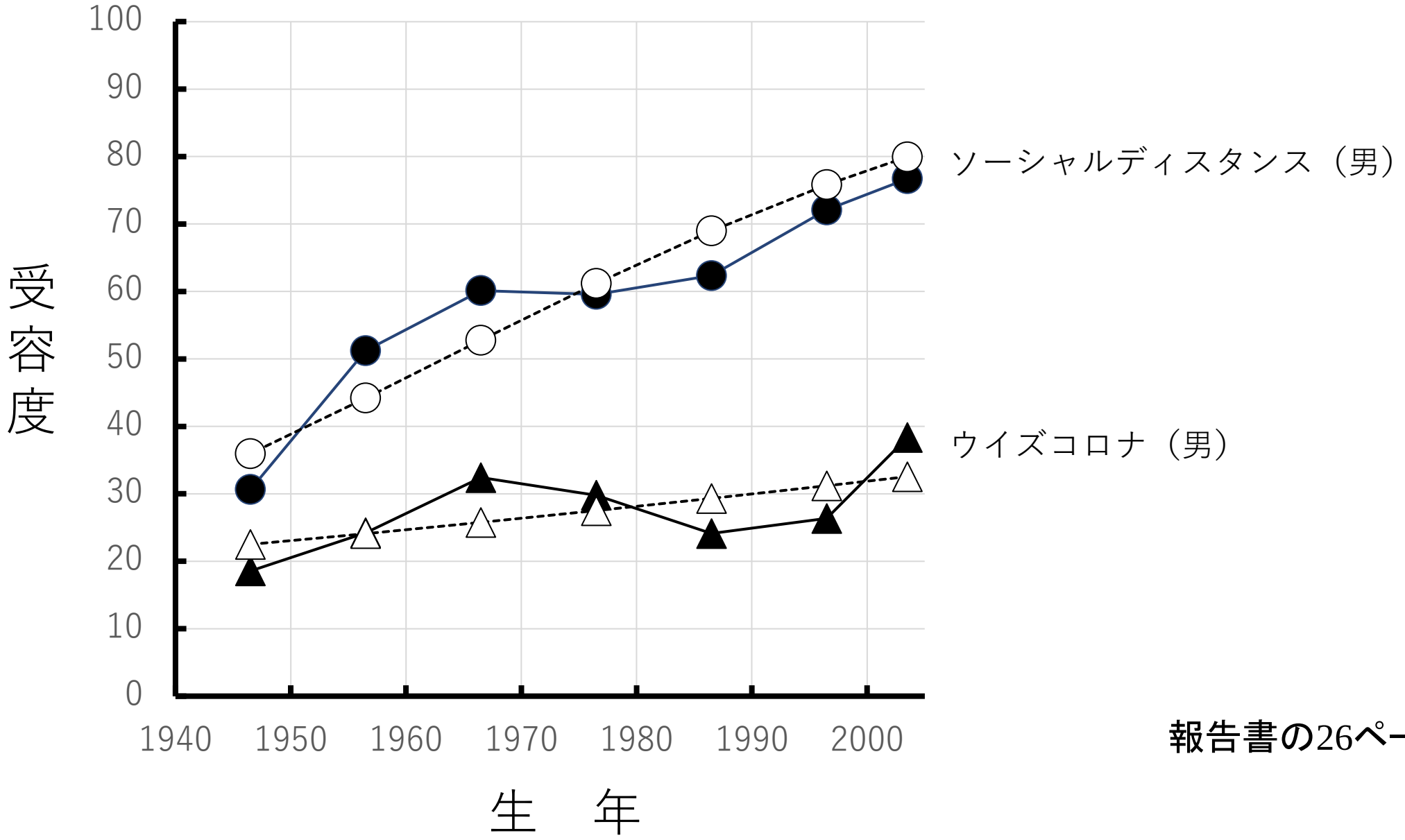
- 次頁から典型的なグラフの例を示す
- 受容度を予測するモデルの的中精度を検討（予測式の掲出は割愛）
- 予測値（点線）と実測値（実線）のフィットを目視で確認


100
90
80
70
60
50
40
30
20
10
0
受容度
1940
1950
1960
1970
1980
1990
2000
生　年
ソーシャルディスタンス（女）
不要不急（男）
ソーシャルディスタンス（男）
ウイズコロナ（男）

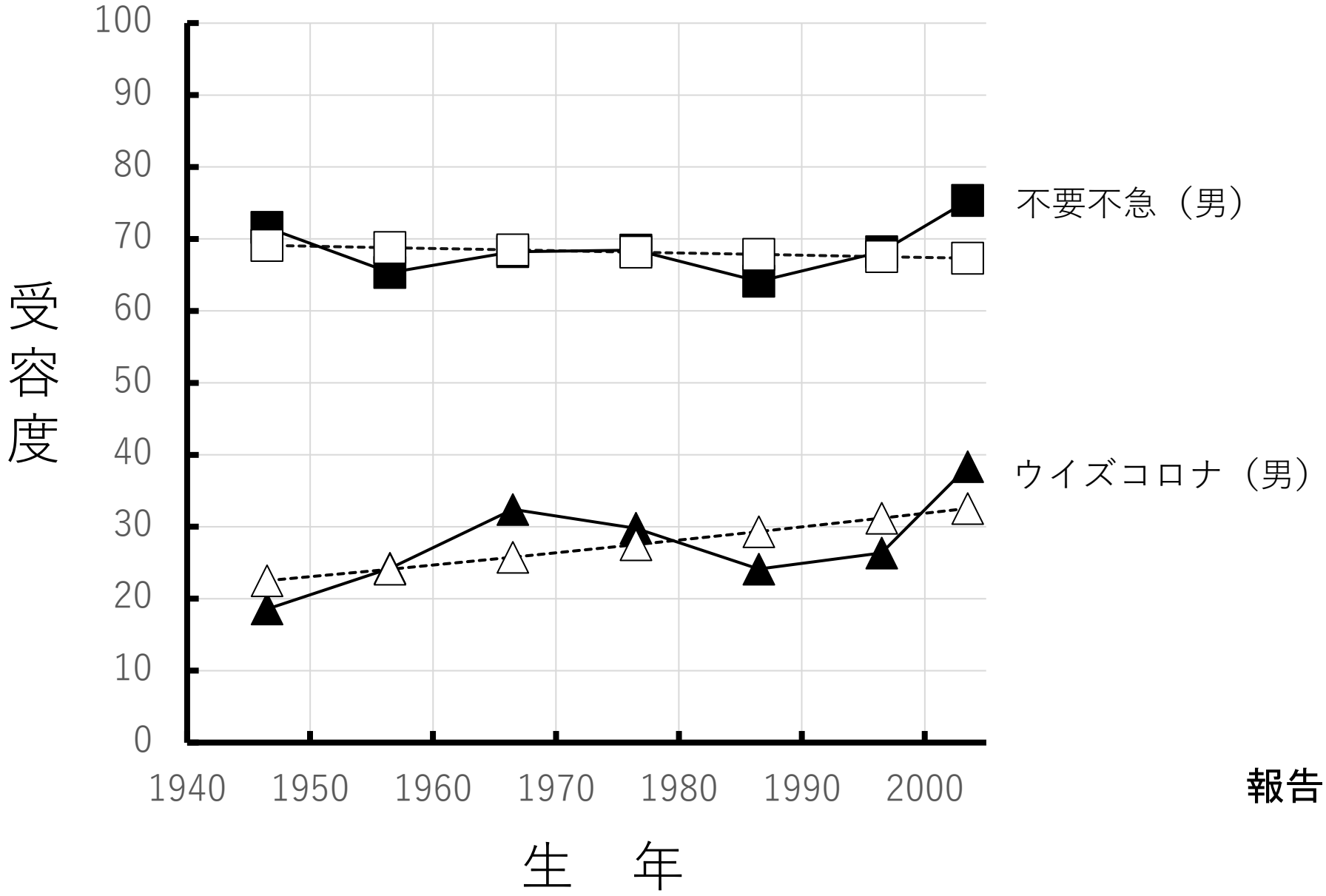
報告書の26ページ


90
80
70
60
50
40
30
20
1940
1950
1960
1970
1980
1990
2000
受容度
生　年
ソーシャルディスタンス（女）
ソーシャルディスタンス（男）

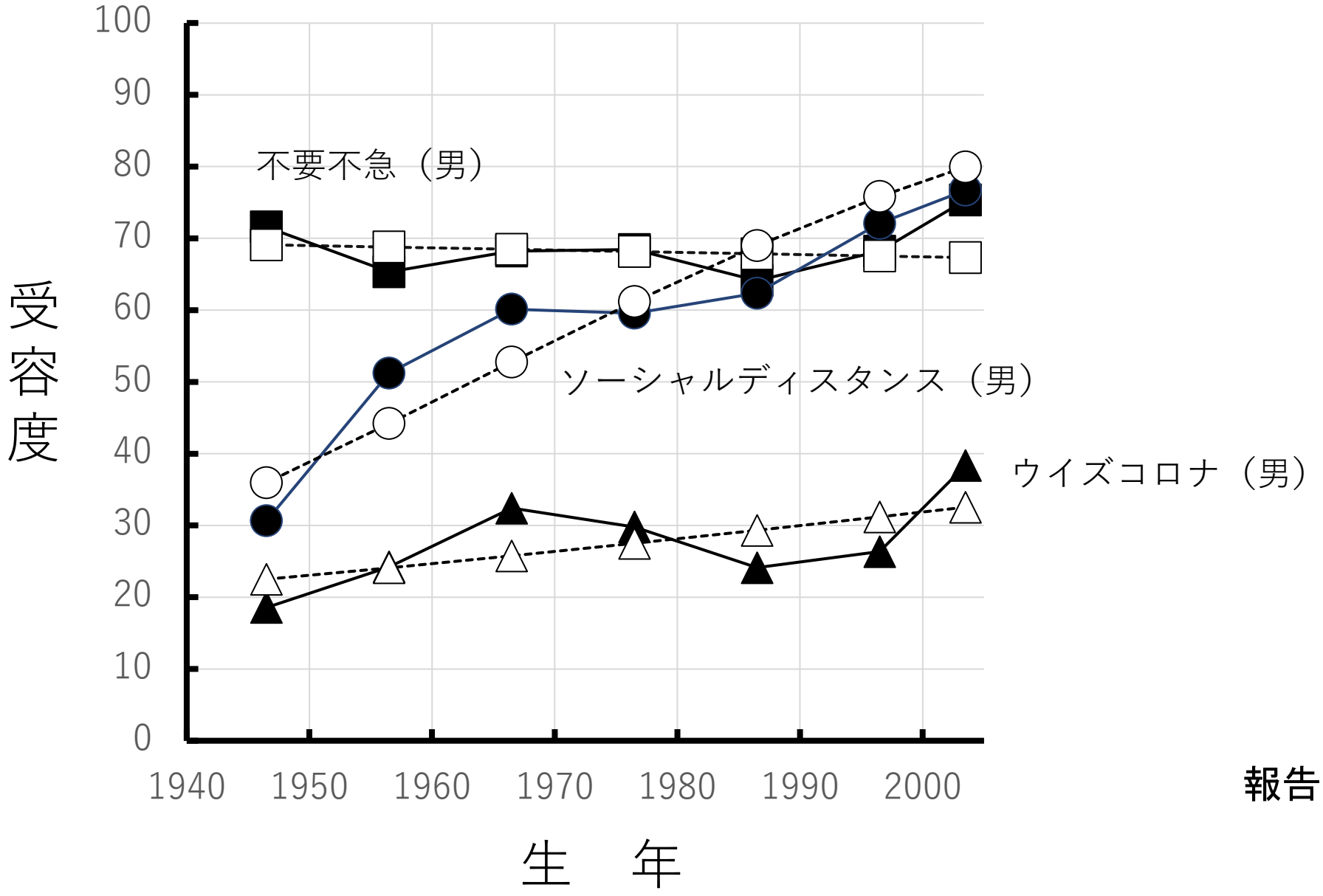
報告書の26ページ


100
90
80
70
60
50
40
30
20
10
0
受容度
1940
1950
1960
1970
1980
1990
2000
生　年
ソーシャルディスタンス（男）
ウイズコロナ（男）

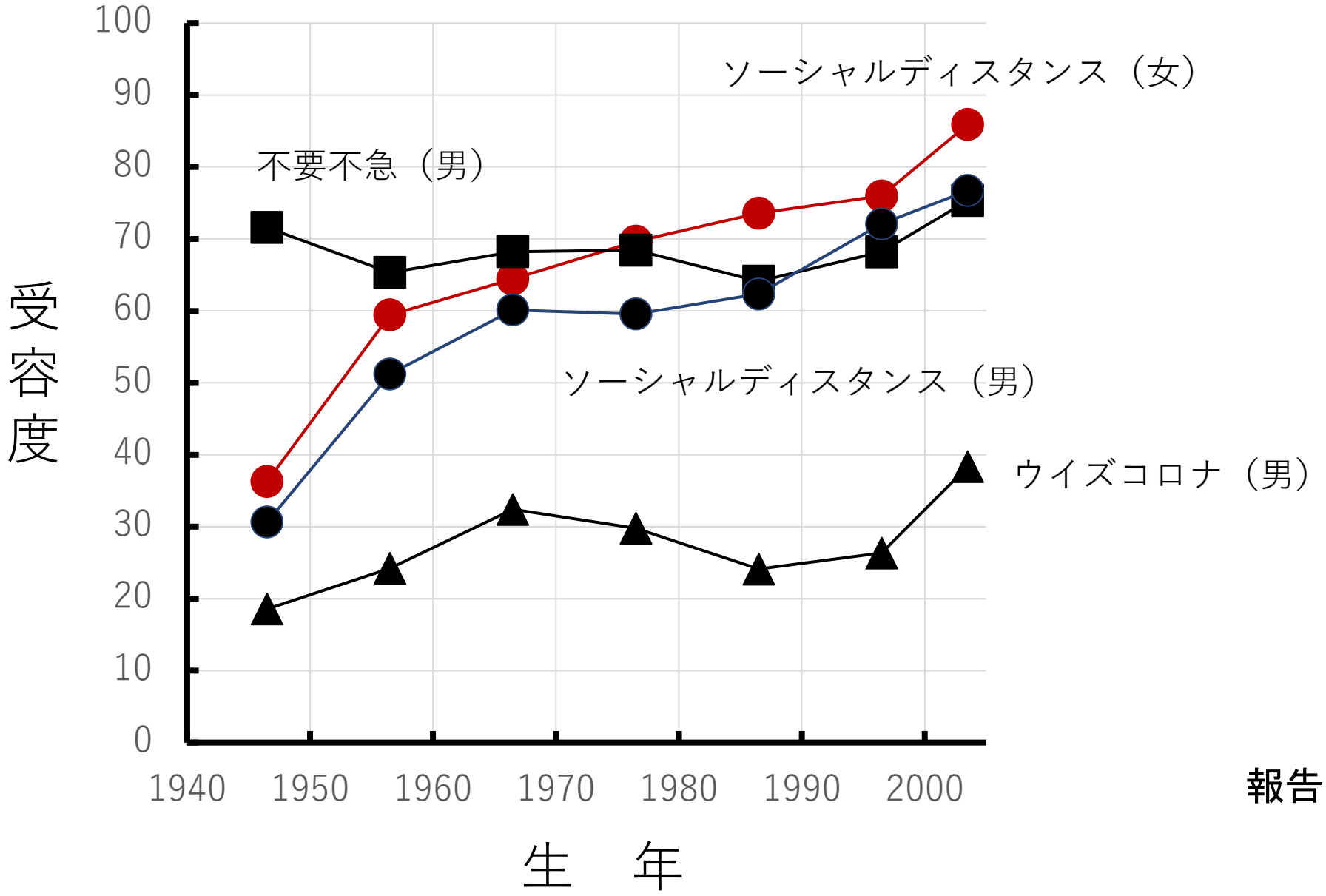
報告書の26ページ


100
90
80
70
60
50
40
30
20
10
0
受容度
1940
1950
1960
1970
1980
1990
2000
生　年
不要不急（男）
ウイズコロナ（男）

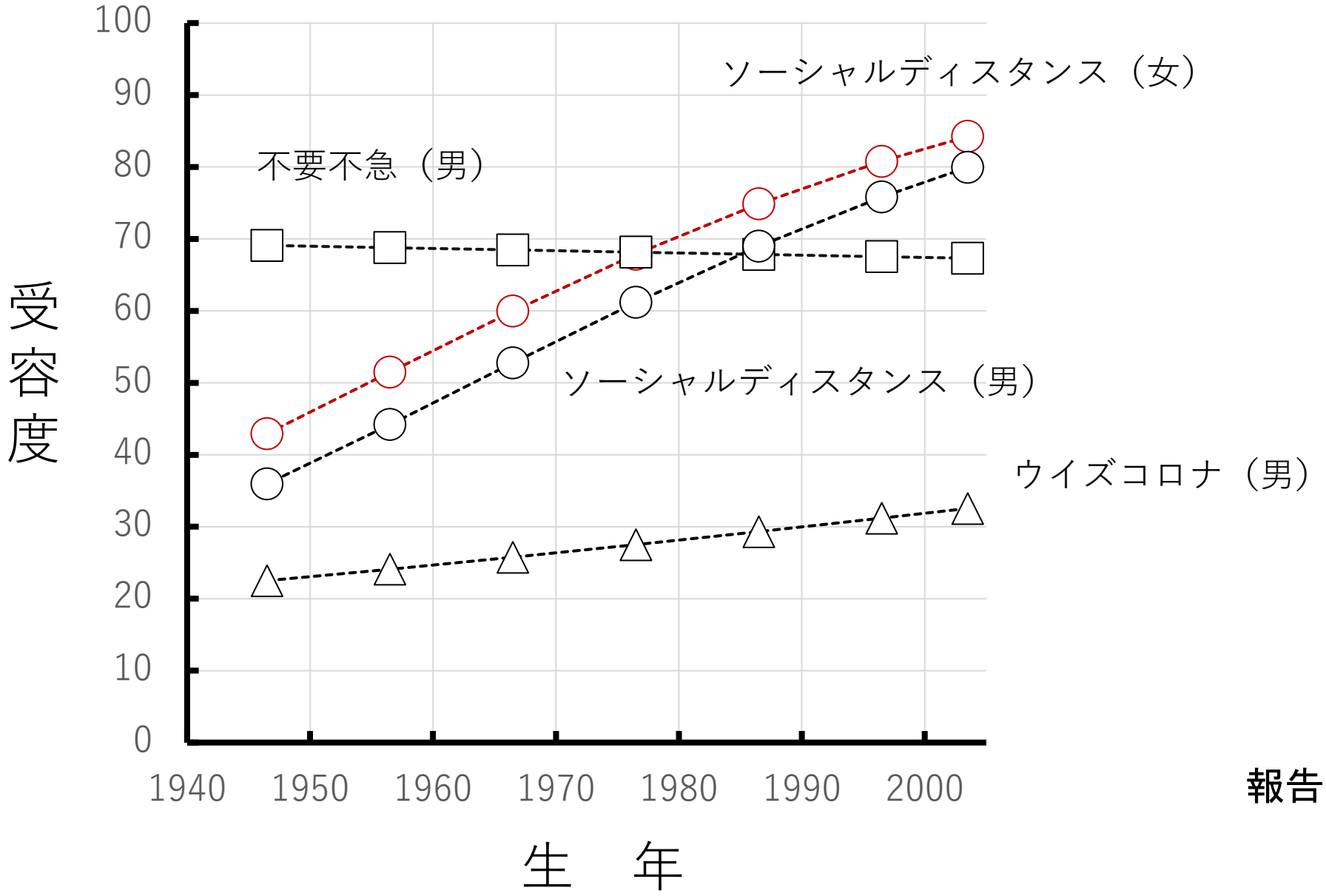
報告書の26ページ


100
90
80
70
60
50
40
30
20
10
0
受容度
不要不急（男）
ソーシャルディスタンス（男）
ウイズコロナ（男）
1940
1950
1960
1970
1980
1990
2000
生　年

報告書の26ページ


100
90
80
70
60
50
40
30
20
10
0
受容度
1940
1950
1960
1970
1980
1990
2000
生　年
ソーシャルディスタンス（女）
不要不急（男）
ソーシャルディスタンス（男）
ウイズコロナ（男）

報告書の26ページ


100
90
80
70
60
50
40
30
20
10
0
受容度
1940
1950
1960
1970
1980
1990
2000
生　年
ソーシャルディスタンス（女）
不要不急（男）
ソーシャルディスタンス（男）
ウイズコロナ（男）

報告書の26ページ

どうする分析

GLMM（一般化線形混合モデル）のロジスティック回帰分析も実行

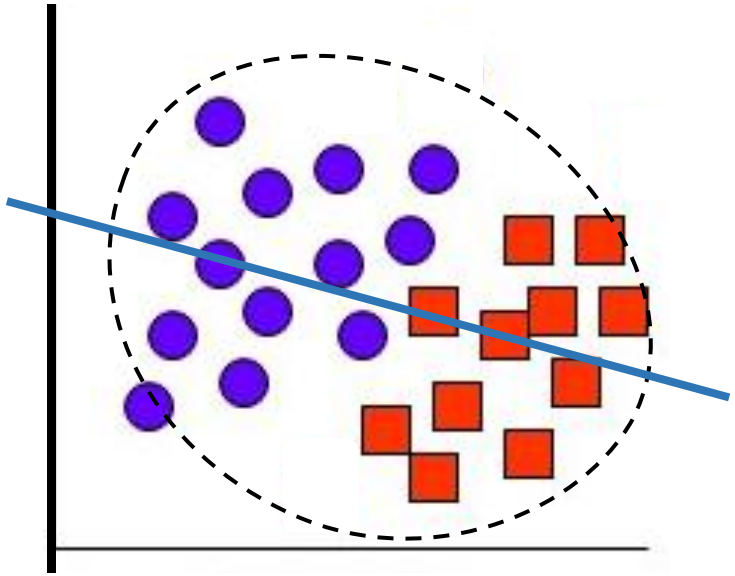
- 8語のデータを一度に投入、語をブロック変数とした
- ランダム傾きモデルやランダム切片モデルなど5種類を設定

- いずれのモデルでもさまざまな警告メッセージが出力される。計算が収束しないという問題のほかに、いろいろあるようにも思えるが詳細は不明
- 不思議なことに、受容度の予測値を出力させるともっともらしい数値が並ぶ
- ところが、出力されたパラメータ（傾きや切片）をもとに以下の数式で検算してグラフ化すると上記の予測値と明らかに合致しないケースがある。不可解！

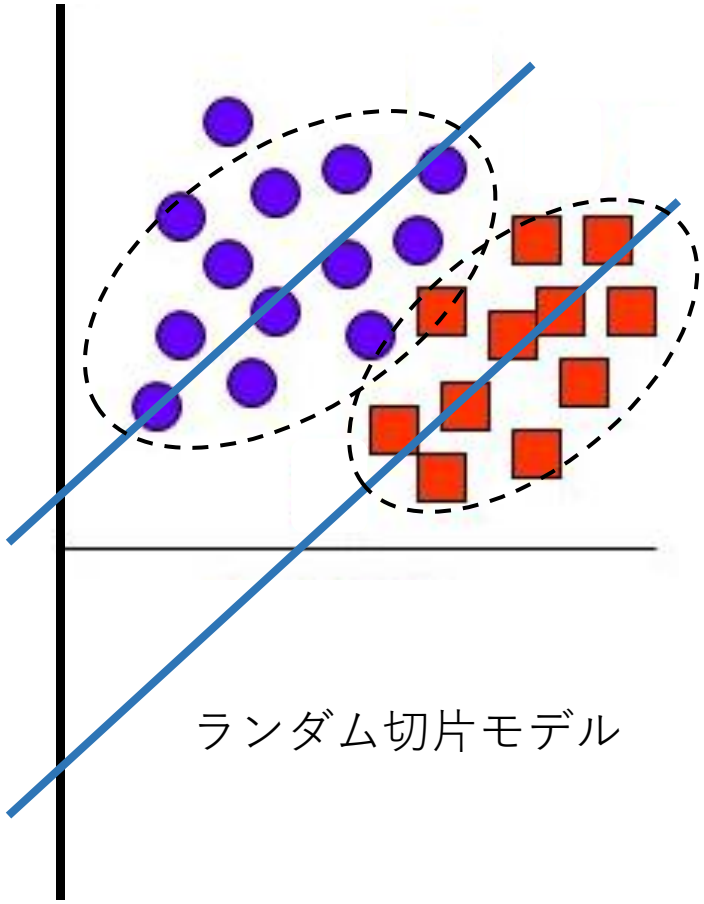
予測確率 $p = 1/[\,1+\exp(-Z)\,]$　ただし $Z = a_1x_1 + a_2x_2 + \ldots + c$

GLMMは生態学的誤謬に対応可能

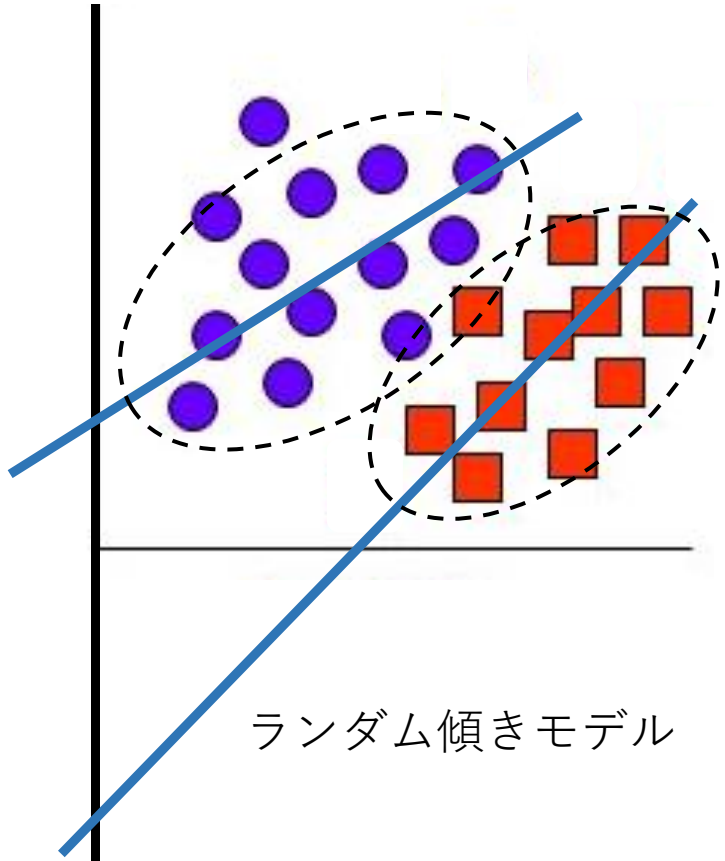
- 生態学的誤謬は中澤（2014）を参照→ http://seisan.server-shared.com/661/661-75.pdf
- すべての人（グループ）を一緒にして分析すると統計モデルの予測精度が低下する
- 個人別（グループ別）に分析すると統計モデルの予測精度が向上する



全体では、有意な相関はない、
あるいは負の相関がある。
ところが、グループ別だとそれ
ぞれ強い正の相関がある。
これを「生態学的誤謬」という



ランダム切片モデル



ランダム傾きモデル

まとめと今後の課題

1. 新語が徐々に世間に広がることにより年代差が生じると考えられている。しかし、今回はどの年代も同時に一気に高頻度で同じ語に接触しているにもかかわらず、カタカナ語で年代差が生じたのはなぜか？
2. 「ソーシャルディスタンス」＝ソーシャル＋ディスタンスで、高年齢層はその要素の語に不慣れなため年代差が大きいのではないか（グラフの傾きが大きい）？
3. ほかに比べて「ウィズコロナ」は世に出るのが遅かったので受容度が低く、年代差も小さい（グラフの傾きが小さい）のではないか？
4. 久屋（2016）が外来語の研究で男女差が明確に出ることを指摘しているが、今回は、カタカナ語すべてについて女性の受容度が男性よりも有意に高いという結果になった。これはなぜか？
5. 「不要不急」はコロナ禍より前からマスメディア等に登場していた（台風接近時など）ので新語とは言えないのでは？
6. 同一年代でも、日常生活で外来語が接触語彙や理解語彙である人と、そうでない人では受容度に差があるのでは？

いずれにせよ、実時間調査すなわち経年調査が必要

参考：社会を対象にしたデータについて

政府機関が公開する統計資料や、久屋（2021）の言う「全国調査のデータ」は、公共財であり、社会インフラである

代表性を有するデータに立脚していることを前提にしている場合が多い

ランダムサンプリングは当然のこととして、さらに

- 年齢層が広い
- 全国規模である

参考：社会を対象にしたデータについて

言語系研究だと

- 日本人の読み書き能力1948年調査：日本における社会調査の原点の一つ
- 国語研による社会言語学の経年・実時間調査（一部に単発的な全国規模調査がある）
- BCCWJなどのコーパス（人間の言語行動の痕跡~~＝残滓の蓄積~~）20230406二重取り消し線
- NHK放送文化研究所の社会調査（放送用語選定の基盤となる「日本語のゆれに関する調査」など）
- 文化庁の国語に関する世論調査

参考：社会を対象にしたデータについて

大学等の研究者が収集するデータのうち代表性を有するものが多い分野は

- 医学：疫学研究などでメディカル・メガバンクを構築
- 経済学：マクロ統計やミクロ統計
- 社会学：社会階層と社会移動全国調査（SSM調査）https://www.l.u-tokyo.ac.jp/2015SSM-PJ/ など
- 統計数理研究所の社会調査：日本人の国民性調査 https://www.ism.ac.jp/kokuminsei/など

おもな引用･参考文献

1. 文化庁「国語に関する世論調査」 https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/kokugo_yoronchosa/index.html
2. 林廷修（2021）「外来語とその言い換え表現からみた言葉の受け入れに関する研究―新型コロナウイルス関連用語を例に―」『日本語学会2021年度春季大会予稿集』61-66
3. 国立国語研究所「外来語」言い換え提案 https://www2.ninjal.ac.jp/gairaigo/
4. 国立国語研究所「病院の言葉」を分かりやすくする提案 https://www2.ninjal.ac.jp/byoin/
5. 久保拓弥（2019）『データ解析のための統計モデリング入門：一般化線形モデル・階層ベイズモデル・MCMC』岩波書店
6. 久屋愛実（2021）「「国語に関する世論調査」に見る外来語の動態―外来語を考える四つの視点―」『日本語学』40(2):84-94, 明治書院
7. 久屋愛実 （2021）「英語由来語彙を公共コミュニケーションでどう運用するべきか：全国調査のデータを活用した福祉言語学的考察」『計量国語学』33巻3号, 130-145 https://www.jstage.jst.go.jp/article/mathling/33/3/33_130/_article/-char/ja/
8. 久屋愛実（2016）「見かけ上の時間を利用した外来語使用意識の通時変化予測」『日本語の研究』12巻4号,69-85 https://www.jstage.jst.go.jp/article/nihongonokenkyu/12/4/12_69/_article/-char/ja/
9. 中澤 渉（2014）「教育データを解釈する－教育社会学における計量分析」『生産と技術』大阪大学生産技術研究会編66(1),75-77 http://seisan.server-shared.com/661/661-75.pdf

【付録】国立国語研究所の研究公正研修「ケーススタディ」（2023年3月9日）で話題になったQRP（Questionable Research Practice：疑わしい研究行為）の防止に向けて：いわゆる「*p*値ハッキング」に関する自己チェックリストの一部分を試作してみました

Q1. 最初に統計的検定をした後、サンプル（データ）を加えて、再度、検定を行いましたか？

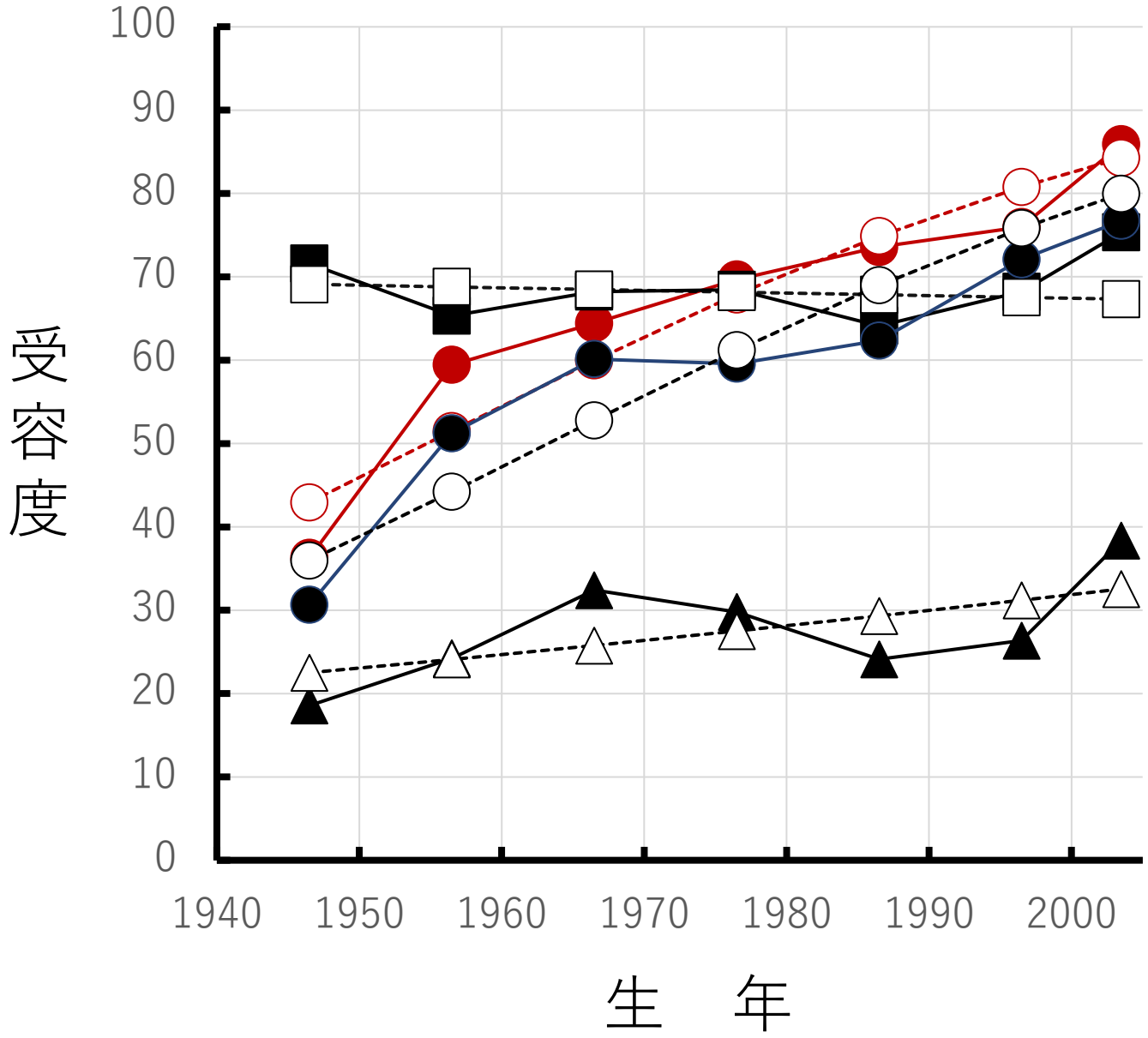
A1. いいえ、最初の1回だけです。国の公開データをそのまま解析しています。

ホンネの感想：大学院生の時代は実験心理学を専攻してました。データの統計的検定でギリギリ有差にならない場合はサンプルを追加するようにと先輩や先生方から教えてもらいました。サンプル数が増えることによって検定力がアップし、有意差が出やすくなるからです。このやり方は、至極当たり前で、実験心理学の研究者なら誰でもやったことがあると思います。ところが、最近になって、それはQRPだとされていることを知り、驚くとともに戸惑ってもいます。

Q2. 外れ値の処理について、どのようなやり方をしたか書いてください

A2. 今回は外れ値の処理は必要ありませんでした。国の公開データをそのまま解析しています。

ホンネの感想：反応時間のデータを統計解析する場合、どうしても外れ値、すなわち異常値のようなものが出現します。これをどういう基準でデータから外すかは、明確なスタンダードがありません。データを出し入れしながら統計的検定を繰り返すことになりますが、これはアウトでしょうか？最初に探索的研究だと宣言しておけば許されるのでしょうか？


100
90
80
70
60
50
40
30
20
10
0
受容度
1940
1950
1960
1970
1980
1990
2000
生　年

報告書の26ページ