独立性制約下の変換の
認知バイアスの補正への適用

神嶌 敏弘[1]，馬場 雪乃[2]，鹿島 久嗣[3]
[1]産業技術総合研究所，[2]筑波大学，[3]京都大学

2020年度人工知能学会全国大会（第34回）@ オンライン，2020-6-10

http://www.kamishima.net

動機

- 公平性配慮型機械学習・データマイニング
  - 社会的にセンシティブな情報 S の影響を排除して予測や変換

↓

- センシティブ情報 S は社会実装のときに排除すべき情報だけに限定されない

- **予測や変換において，その影響を排除したい情報の削除に利用可能**

↓

- **人間の認知バイアスの除去への効果を調べる**

形式的公平性

機械学習の公平性では次の影響を考慮する

センシティブ特徴 S → **影響** → **結果・目的** Y

- 社会的にセンシティブな情報
- 法令・規則で制限された情報
- その他無視すべき情報

- 大学入試
- 与信スコア
- 広告クリック率

形式的公平性

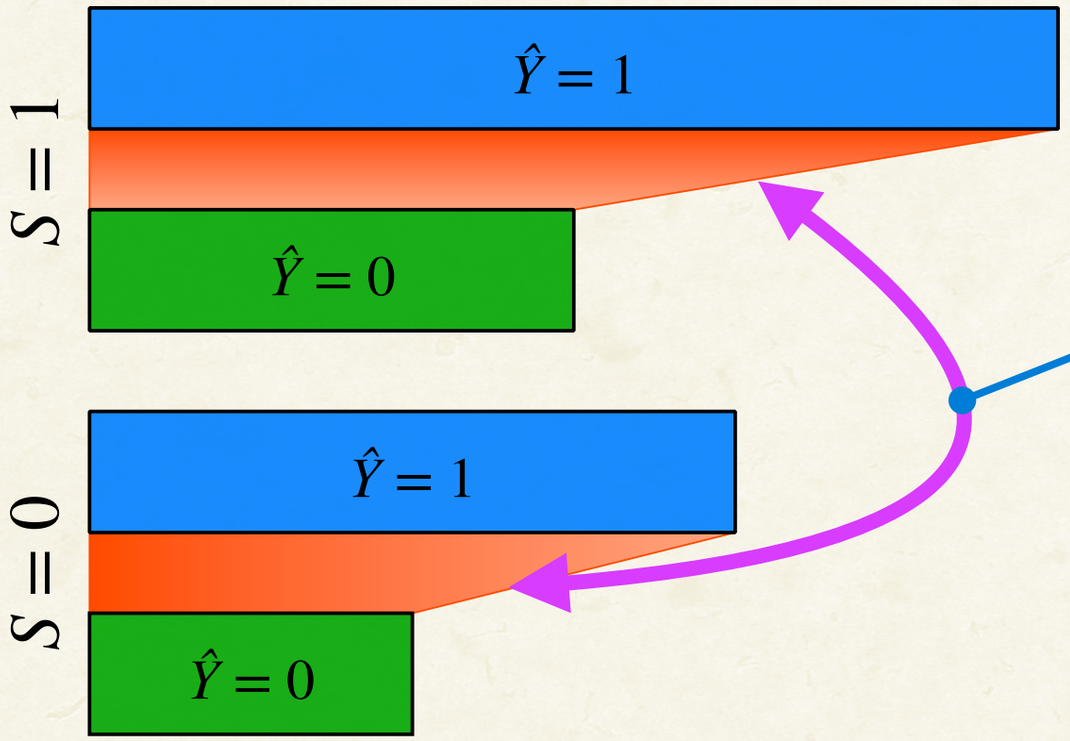
モデル中のセンシティブ特徴 S, その他の特徴 $\mathbf{X}$, 目的変数 Y の間の形式的な関係で定義されるある望ましい状態

- どのような関係を考えるか
- どの変数集合の関係を考えるのか
- センシティブ変数や目的変数のどの状態を考慮するのか

独立性 / 統計的均一性

配分の公正 ➡ 独立性 (independence)
統計的均一性 (statistical parity)

$$\hat{Y} \perp\!\!\!\perp S$$



センシティブ情報の値によらず,
予測結果の値の比率は一定

センシティブ情報 S は予測結果 $\hat{Y}$ に

直接的だけでなく，間接的にも影響を与えない

認知バイアス

[Eickhoff 18]

- **認知バイアス**：利用者の応答は，入力インターフェースなどの要因によって偏りを生じる
  - 限定合理性やナッジなどとも関連
- **曖昧性効果** (ambiguity effect)：情報が欠落していて決定が難しいときには，選ばれにくい
- **アンカリング** (anchoring / focalism)：（最初に見せてしまうなど）特定の情報に不均一に注目してしまう
- **バンドワゴン効果** (Bandwagon effect)：グループの行動に従って判断する
- **おとり効果** (decoy effect)：選択肢 A と Bがあるときに，B と似ているが明らかに劣った C を見せることで，B を好むようにさせる

寿司の一対比較実験

10種類の寿司

トロ　マグロ　エビ　イクラ　アナゴ　ウニ　テッカ　イカ　タマゴ　カッパ

人気 ⟵ ⟶ 不人気

[Kamishima　2003]　の5000人のデータでは左の方が人気で，この順位を使って認知バイアスの検証を行う

※ 今回データではエビが凋落しており6位になっていたが他は同じ順序

寿司の一対比較実験

クラウドソーシングを用いていろいろな条件下で質問する

通常の質問

- 左右に寿司を並べ好きな方を選択
- 一人当たり48対の比較データを収集

集中度テスト

- 50質問のうち2件は集中度テスト
- 両方のテストで右を選んだデータのみを採用
- どの条件でも100件前後のデータを収集

寿司の一対比較実験

以下の条件で収集したデータを分析

- **ベースライン**：左右に寿司を無作為に割り付けた
  - **実験研究によって $\hat{Y} \perp\!\!\!\perp S$ の条件を達成する**
- **バンドワゴン効果・人気順**：人気アイテムを強調
- **バンドワゴン効果・不人気順**：不人気アイテムを偽って強調
- **位置バイアス・人気順**：人気のある寿司は必ず左に表示する
- **位置バイアス・不人気順**：人気のある寿司は必ず右に表示する

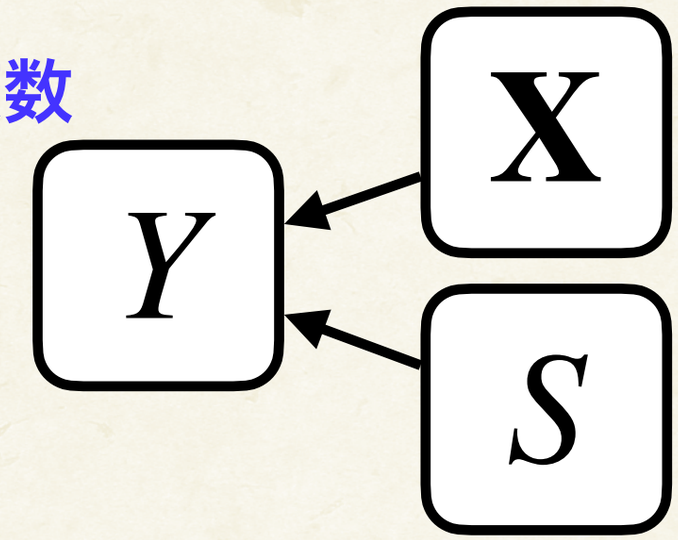
バンドワゴン効果効果のための強調の例

問題設定

非センシティブ変数

目的変数

寿司 i と寿司 j のうち
どちらが好みかの決定



比較する寿司 i と寿司 j を指定

認知バイアスで寿司 i と寿司 j のどちらが有利か

センシティブ変数

$$\Pr[\, Y \mid S, \mathbf{X} \,]$$

寿司 i と寿司 j が $\mathbf{X}$ で，センシティブ変数 S で認知バイアスが与えられたとき，寿司 i を寿司 j より好む確率を全ての (i, j) について推定

モデル

Babington Smithモデル：一対比較の飽和モデル

S の各値ごとに，寿司 i が寿司 j より好まれる割合を求める

汎化しないので，未観測の場合には値が不定

↓

Y の分布を変数 S によらない値（算術平均）にして消す

$$\tilde{\Pr}[Y \mid \mathbf{X}] = 0.5 \Pr[Y \mid S = 0, \mathbf{X}] + 0.5 \Pr[Y \mid S = 1, \mathbf{X}]$$

↓

一部の (i, j) についてはデータが観測されない

※ 人気アイテムを認知バイアスで有利にすると，寿司 j が i より不人気の場合にはデータが観測されない

↓

観測されない事象については $\Pr[Y \mid S, X] = 1/2$ として計算

因果効果との関係

いずれか一方しか観測できず，もう一方は反実仮想になる

$S=0$：有利なバイアス	➡	Y_1：有利なバイアス下で寿司 i を選択
$S=1$：不利なバイアス	➡	Y_2：不利なバイアス下で寿司 i を選択

因果推論：因果効果の推定

$$\mathrm{E}[Y_1] - \mathrm{E}[Y_2]$$

有利なバイアスのために，どれくらい寿司 i は選択されやすくなるか

↕

S に影響されない Y

$$Y = Y_1 \Pr[S = 0] + Y_2 \Pr[S = 1]$$

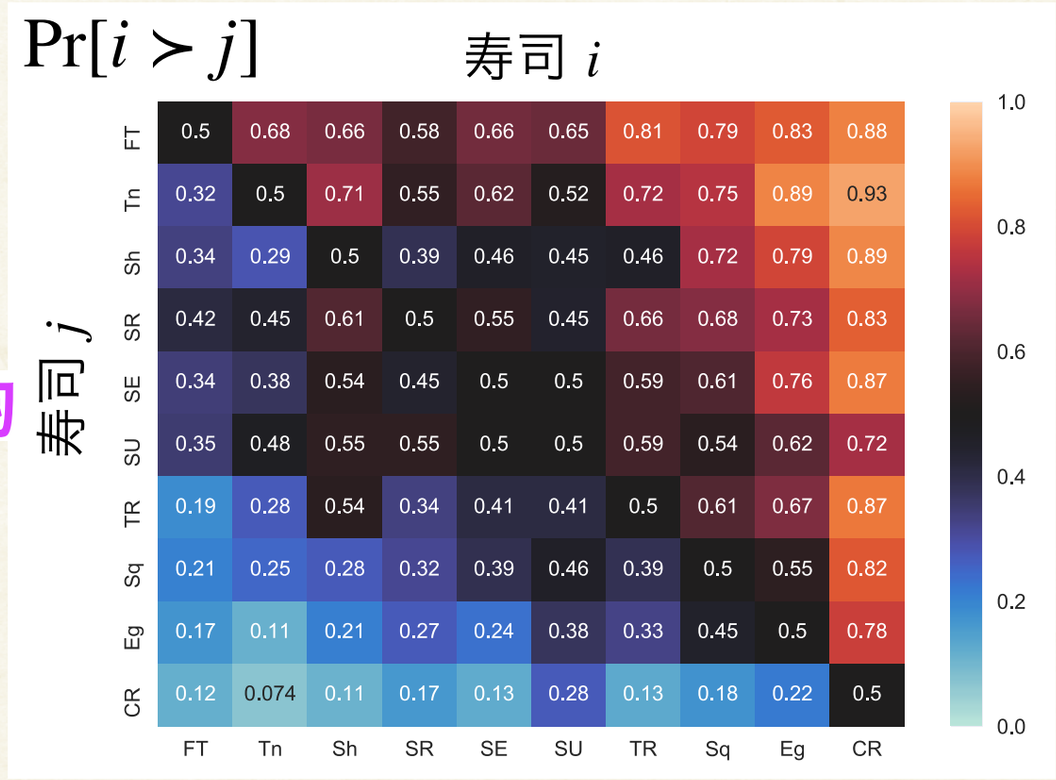
混合モデルで表される変数に注目

実験

認知バイアスのあるデータの寿司の人気度から，無作為割り付けをしたベースラインの寿司の人気度を引いた差を求める

寿司 i の人気度

$\Pr[i \succ j]$ を示した行列の列方向の平均

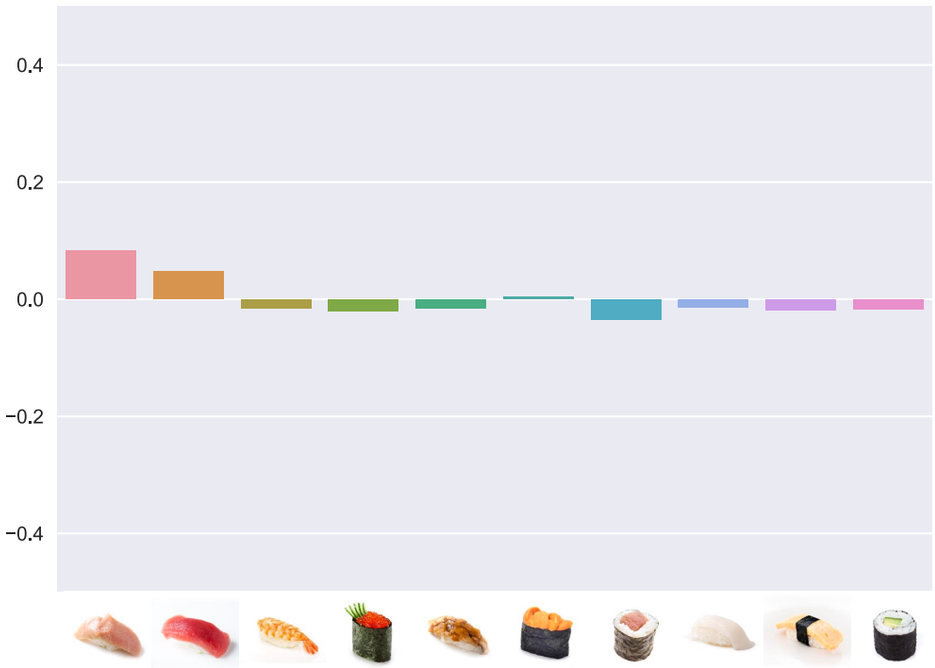


- **補正前**：$\Pr[Y \mid \mathbf{X}] = \Pr[Y \mid S=0, \mathbf{X}] \Pr[S=0 \mid \mathbf{X}] + \Pr[Y \mid S=1, \mathbf{X}] \Pr[S=1 \mid \mathbf{X}]$
- **補正後**：$\tilde{\Pr}[Y \mid \mathbf{X}] = 0.5 \Pr[Y \mid S=0, \mathbf{X}] + 0.5 \Pr[Y \mid S=1, \mathbf{X}]$

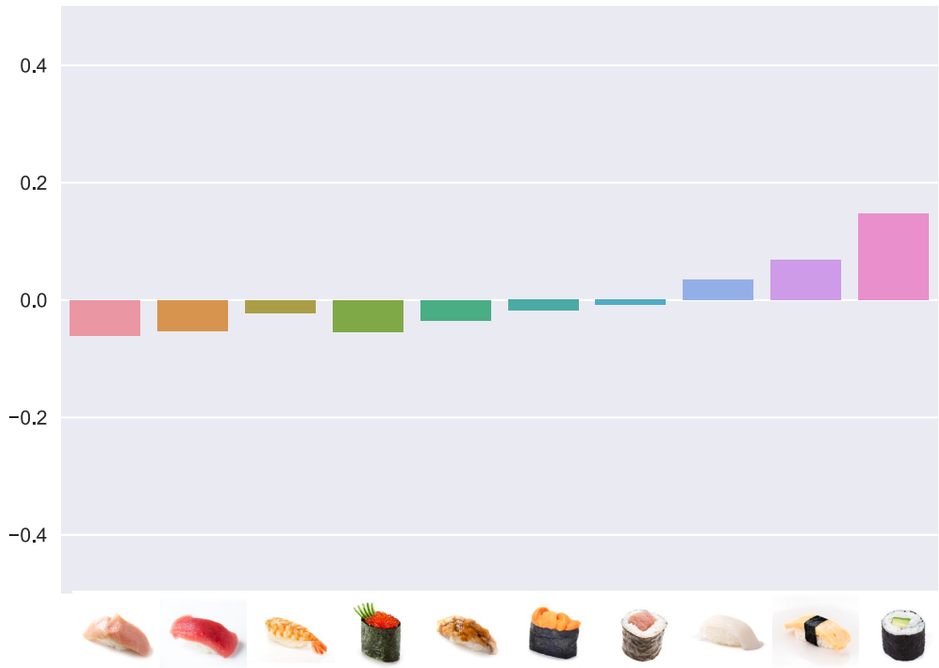
実験結果（バンドワゴン効果）

バンドワゴン効果：人気アイテムを強調

補正前： $\Pr[Y \mid \mathbf{X}]$



補正後： $\tilde{\Pr}[Y \mid \mathbf{X}]$

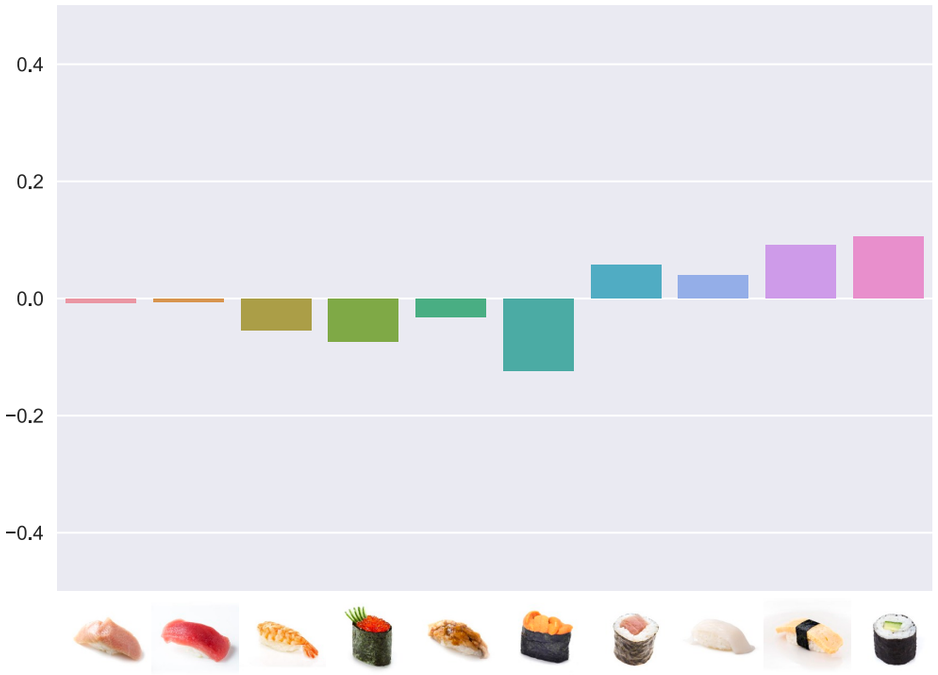


- 人気アイテムを強調しても，あまり差はでない
- むしろ補正後の方が差が顕著に

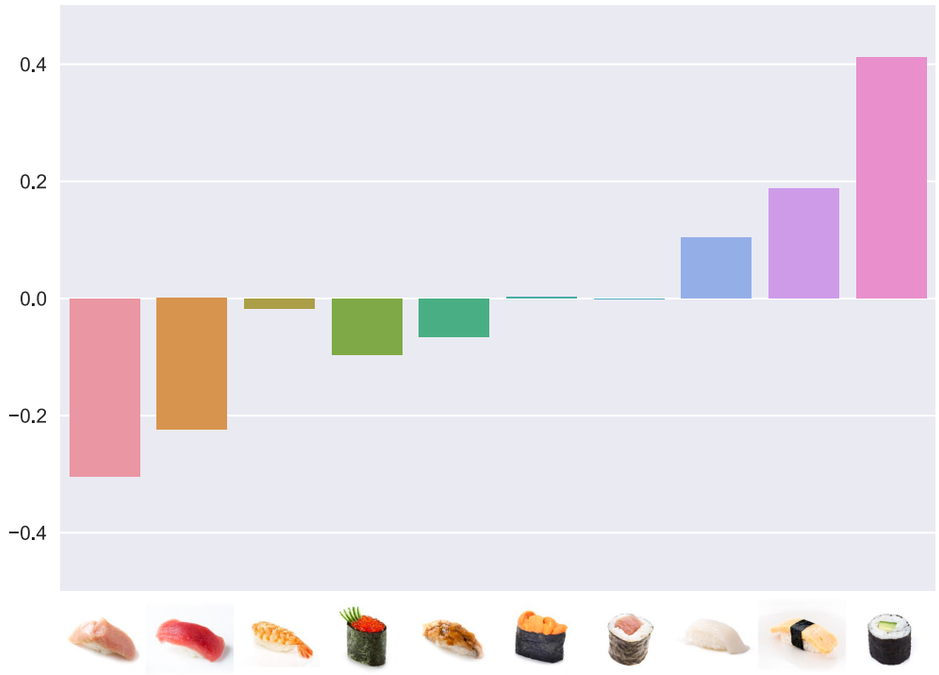
実験結果（バンドワゴン効果）

バンドワゴン効果：不人気アイテムを強調

無正前： $\Pr[Y \mid \mathbf{X}]$



補正後： $\tilde{\Pr}[Y \mid \mathbf{X}]$

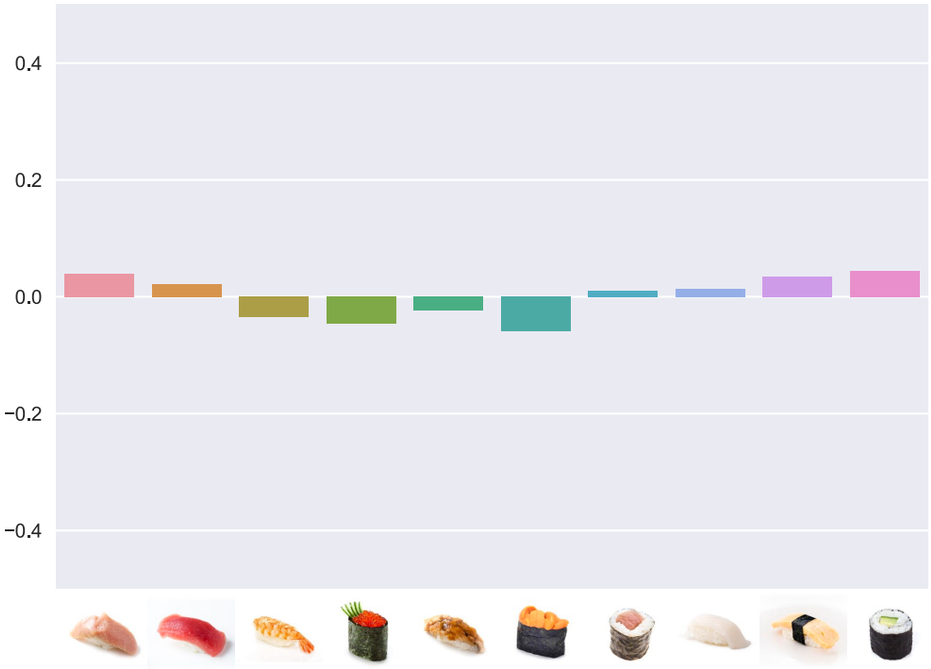


- 人気アイテムを強調するより変化は大きい
- やはり補正後の方が差が顕著に

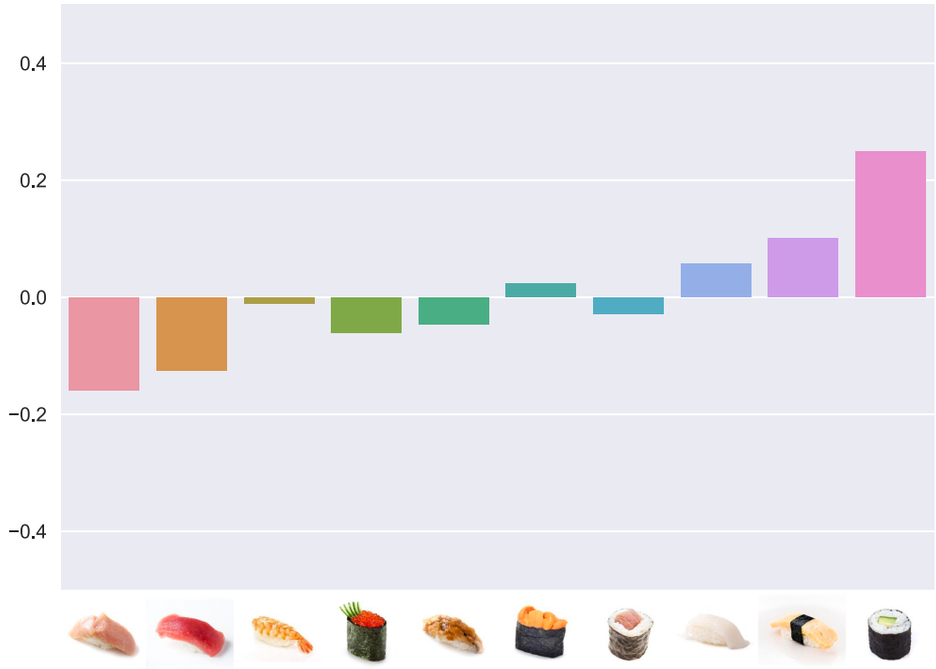
実験結果（バンドワゴン効果）

バンドワゴン効果：人気と不人気の併合データ

無正前： $\Pr[Y \mid \mathbf{X}]$



補正後： $\tilde{\Pr}[Y \mid \mathbf{X}]$

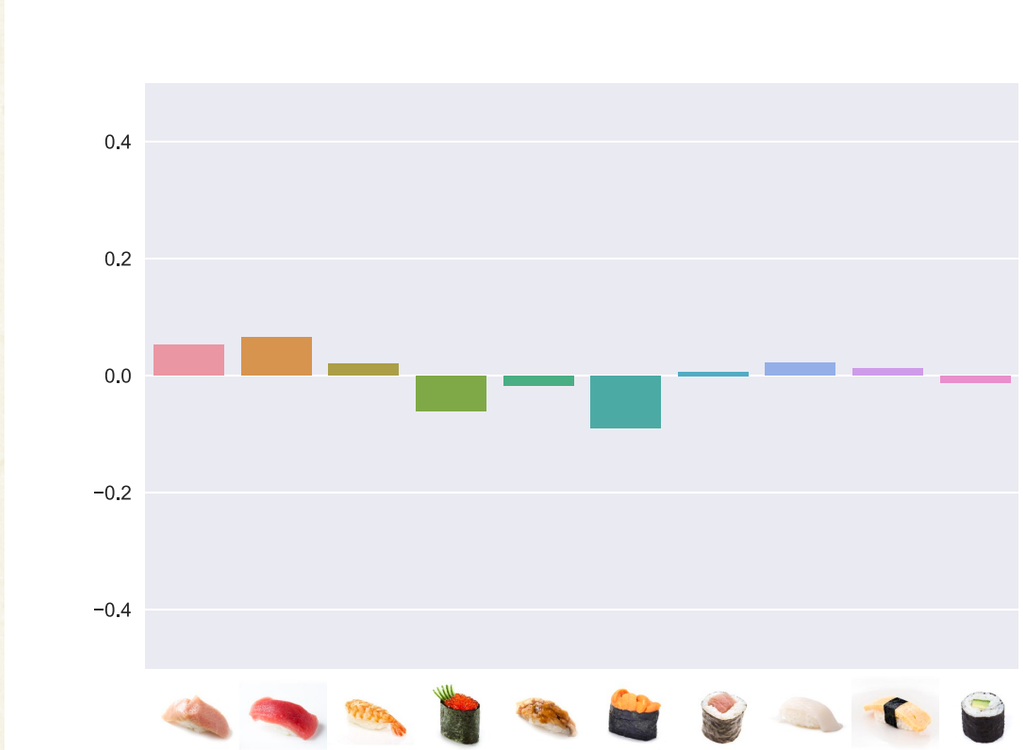


- 両者を足しても打ち消し合う分けではない
- やはり補正後の方が差が顕著に

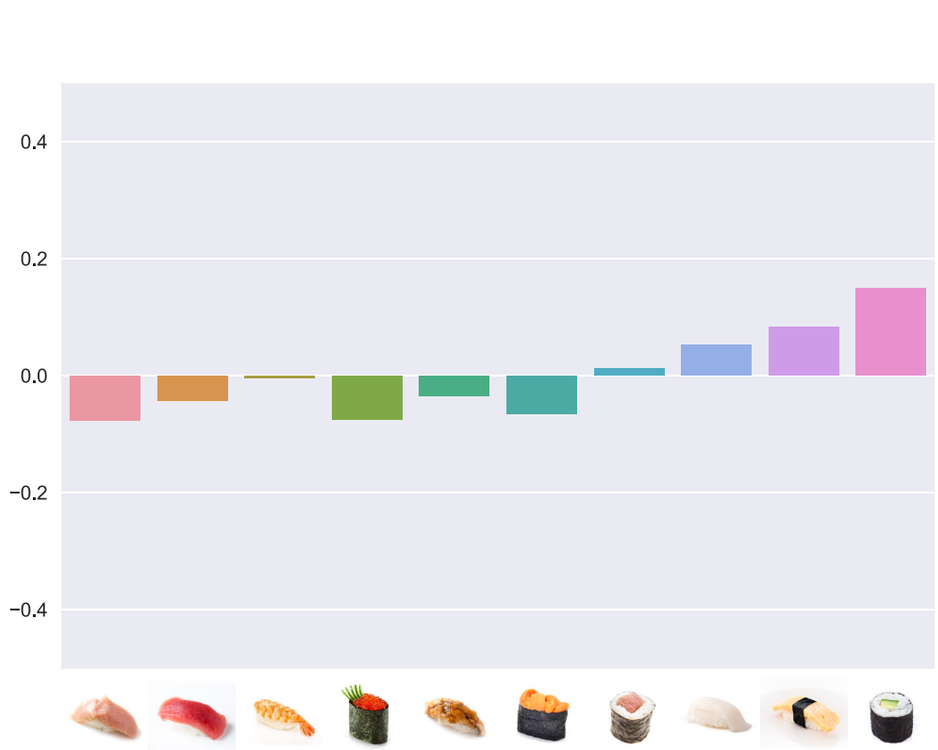
実験結果（位置バイアス）

位置バイアス：人気アイテムが左

無正前： $\Pr[Y \mid \mathbf{X}]$



補正後： $\tilde{\Pr}[Y \mid \mathbf{X}]$

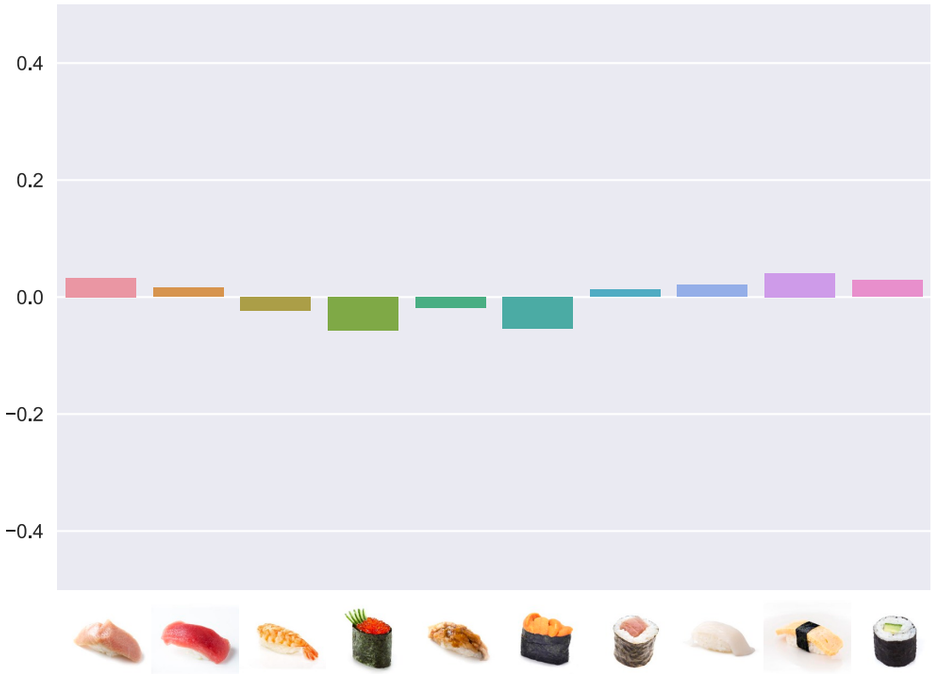


- 人気アイテムがより選ばれ易くなるわけではない
- やはり補正後の方が差が顕著に

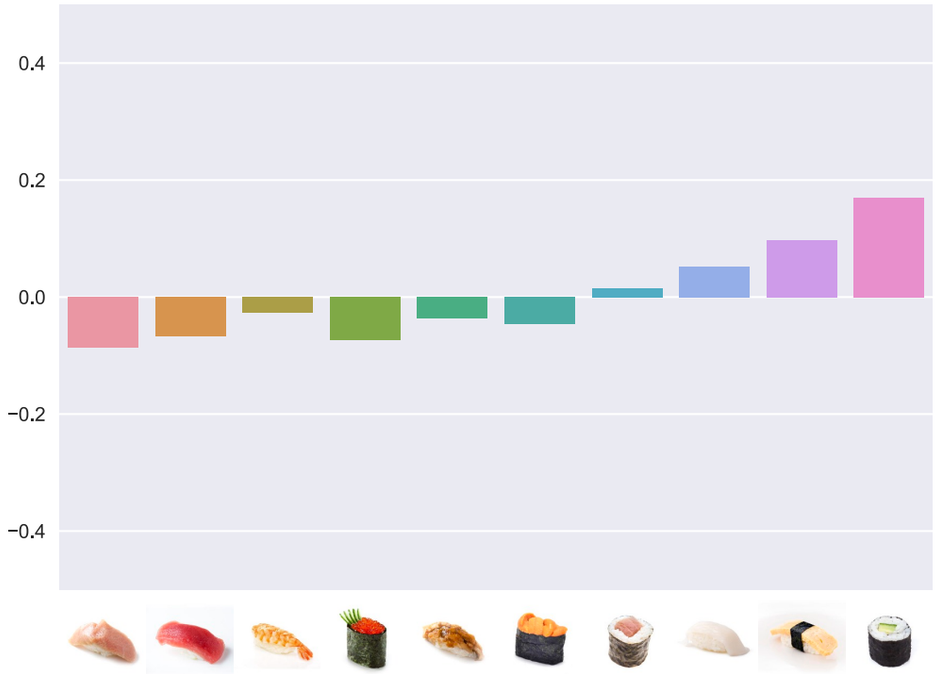
実験結果（位置バイアス）

位置バイアス：不人気アイテムが左

無正前： $\Pr[Y \mid \mathbf{X}]$



補正後： $\tilde{\Pr}[Y \mid \mathbf{X}]$

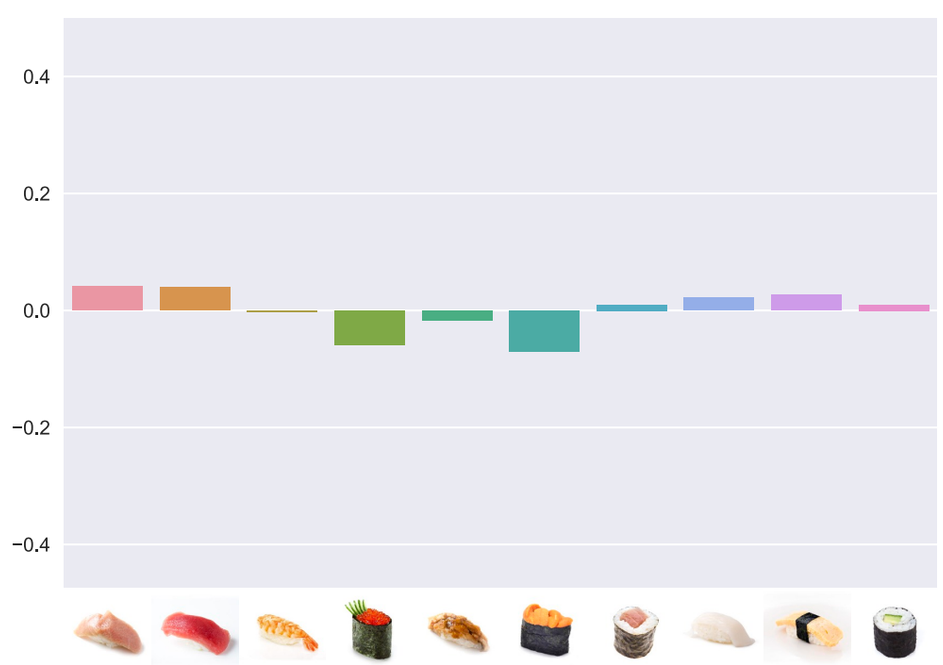


- 不人気アイテムがより選ばれ易くなるわけではない
- やはり補正後の方が差が顕著に

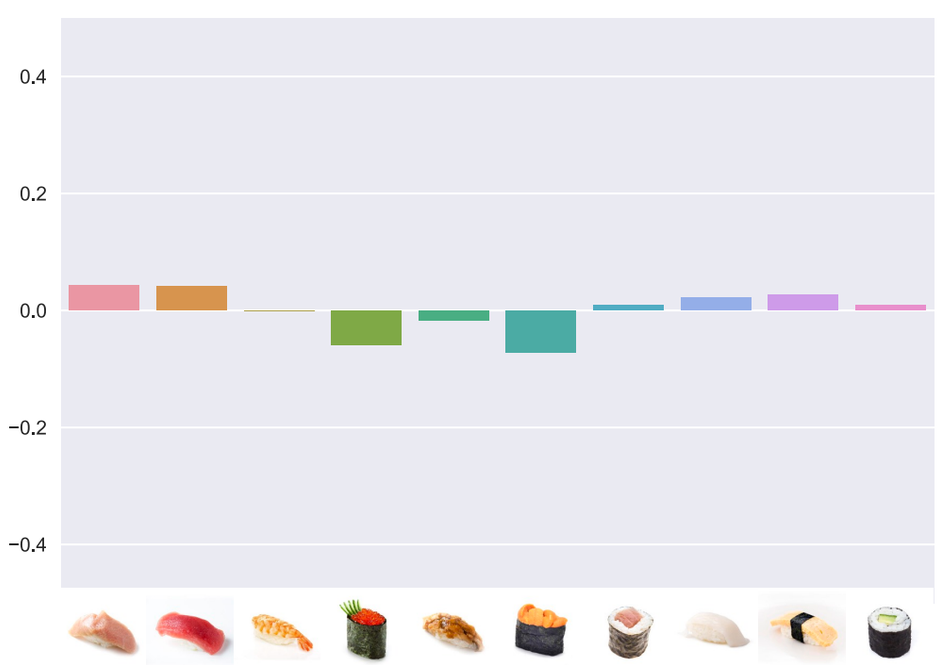
実験結果（位置バイアス）

位置バイアス：人気と不人気の併合データ

無正前： $\Pr[Y \mid \mathbf{X}]$



補正後： $\tilde{\Pr}[Y \mid \mathbf{X}]$



- 併合データは無作為割り付けと変わらないはずだが，やはり差は残る
- S の分布はほとんど同じなので，ほとんど変わらない

まとめ

まとめ

- 認知バイアスの影響を，観察研究で除外するための補正方法を検討した
- 被験者実験で無作為割り付けをしたデータと比較したが，補正の効果は確認できなかった

今後の予定

- 未観測データの補完手法が問題と思われるため，汎化のできるBradley-TerryやMallowsなどのモデルを $\mathrm{P}[Y \mid S, \mathbf{X}]$ の記述に導入する

※ 本研究はJSPS科研費JP24500194，JP15K00327，およびJP18H03300の助成を受けた．