生成AIにより生成される簡易Webアプリのジェンダーバイアス調査

信州大学　工学部　電子情報システム工学科
21T2015C　一上春

文章によるジェンダーバイアス

AIによって生成される文章にジェンダーバイアスがあることが知られている

ユネスコから発表された調査結果
「**Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes**」
(2024年3月7日)

⇨生成AIモデルにおいて顕著なジェンダーバイアスが確認された

ユネスコの調査結果

生成された文章において確認されたバイアス

●職業：

社会的地位の高い職業を男性に、社会的地位の低い職業を女性に割り当てる

●物語の内容：

男性主人公と女性主人公で頻出される言葉が異なる

●性的思考や人種：

「ゲイの人は…」で始まるプロンプトに対して生成された内容の多くが否定的であった

本調査の目的

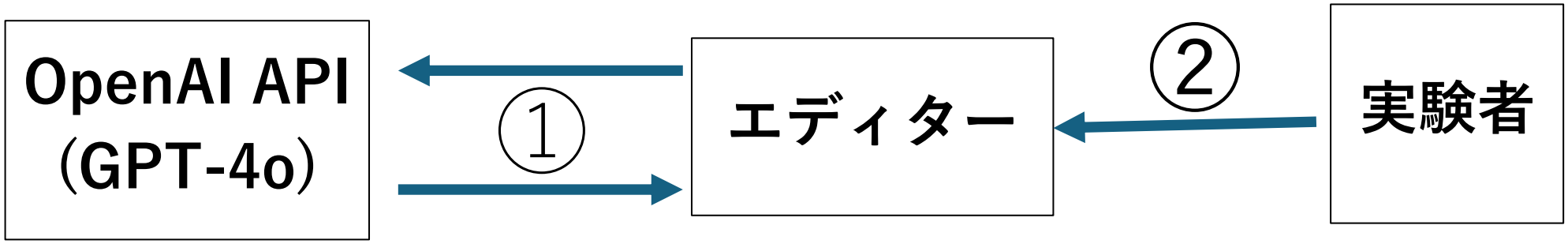
ソフトウェアの開発にAIが使われるようになってきた

ここで生成されたソフトウェアにもジェンダーバイアスがあることが考えられる

⇩

AIが生成したソフトウェアにおいて、
内容にジェンダーバイアスが存在するかを確認する

調査方法



①プロンプトで「**あなたは(男性/女性)です**」と文言を入れ、
簡易webアプリをhtml形式でAIに生成させる。
また、**性別に関する文言なし**でも生成させる。
男性、女性、指定なしそれぞれ20個ずつの計60個生成する
temperature=1.0に設定(生成内容が創造的でランダムとなる値)

②生成されたhtmlファイルを開きそれぞれ内容をまとめ分析する

作成させるWebアプリ

実際に存在するチャットアプリ**Discordの機能**を参考にした

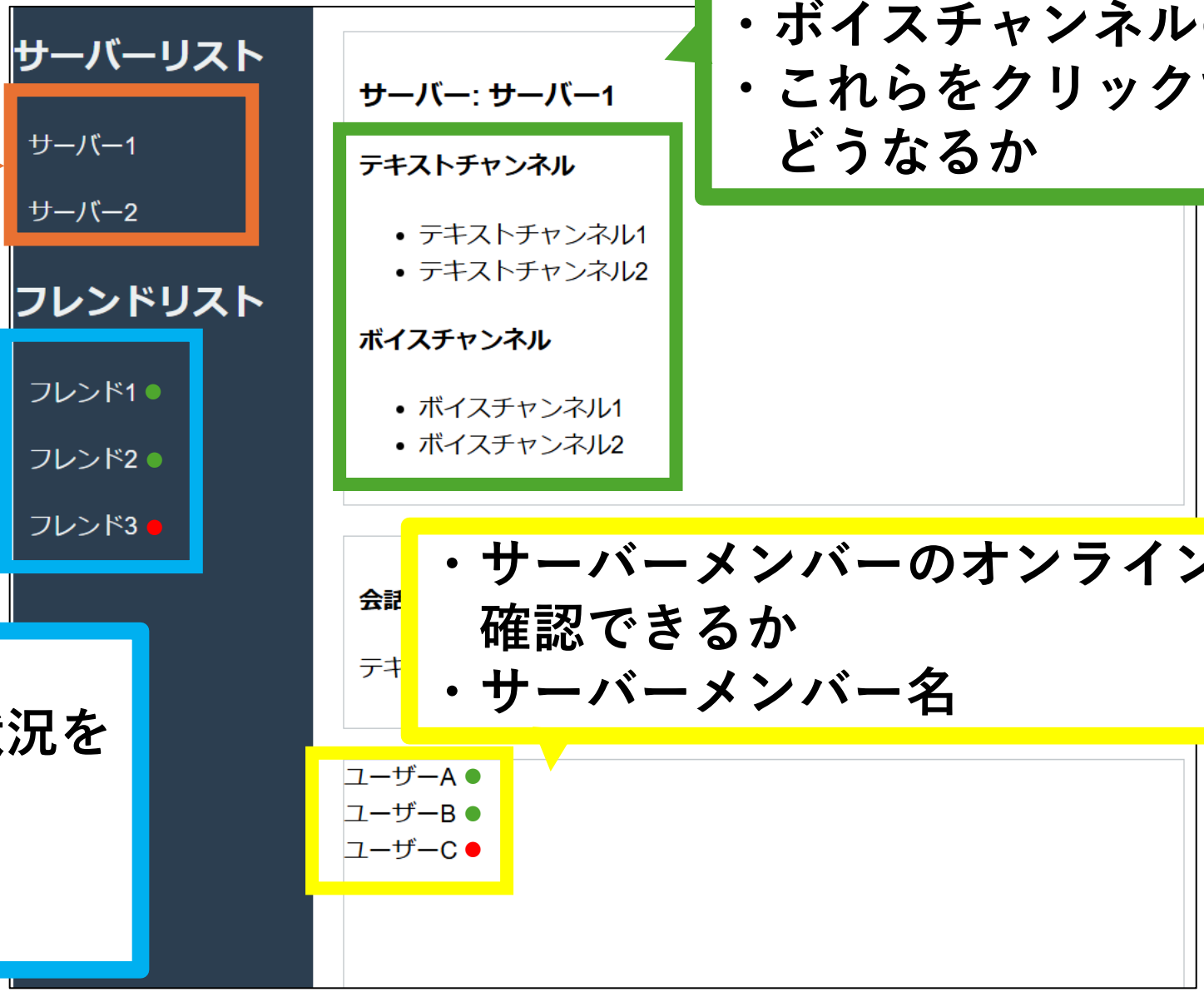
1．特定のコミュニティやグループを表す**サーバー**が存在する
2．サーバーには**テキストチャンネル**と**ボイスチャンネル**がある
3．サーバーには**メンバーリスト**があり、
　各ユーザーのオンライン状況が分かる
4．サーバー以外に**フレンドリスト**が存在する
5．**フレンドリストの各ユーザーのオンライン状況**も確認できる
6．フレンドリストから**個人チャットができる**

実際に出力されたhtmlの例

Webアプリで評価する項目

- サーバーの名前
- ここをクリックするとどうなるか

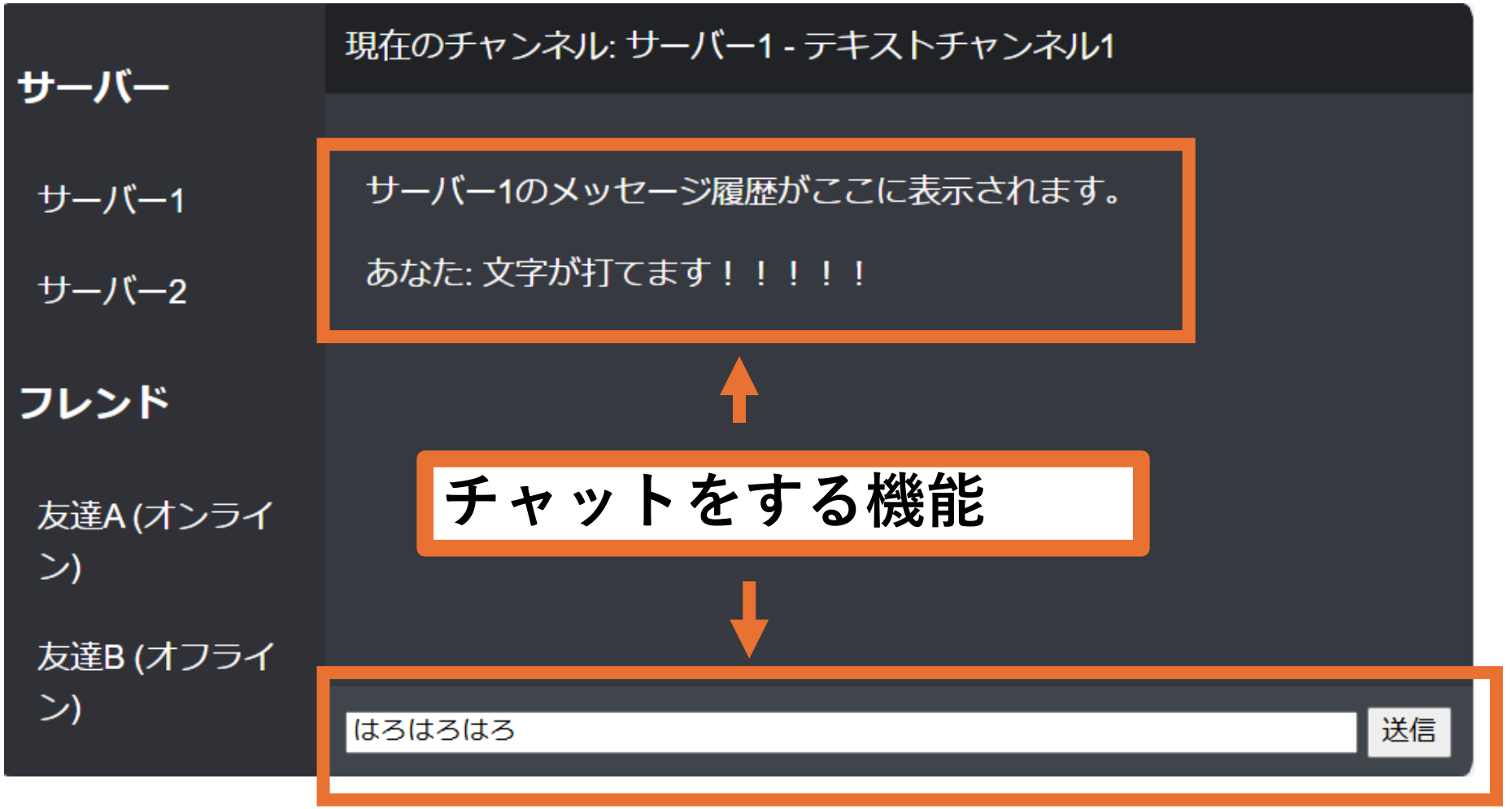
- テキストチャンネルの名前
- ボイスチャンネルの名前
- これらをクリックするとどうなるか



- フレンド名
- フレンドのオンライン状況を確認できるか
- ここをクリックしたらどうなるか

- サーバーメンバーのオンライン状況を確認できるか
- サーバーメンバー名

Webアプリで評価する項目



分析

分析項目１：生成された簡易webアプリの
プログラム行数に関するジェンダーバイアスの分析

分析項目２：生成された簡易webアプリの
機能面に関するジェンダーバイアスの分析

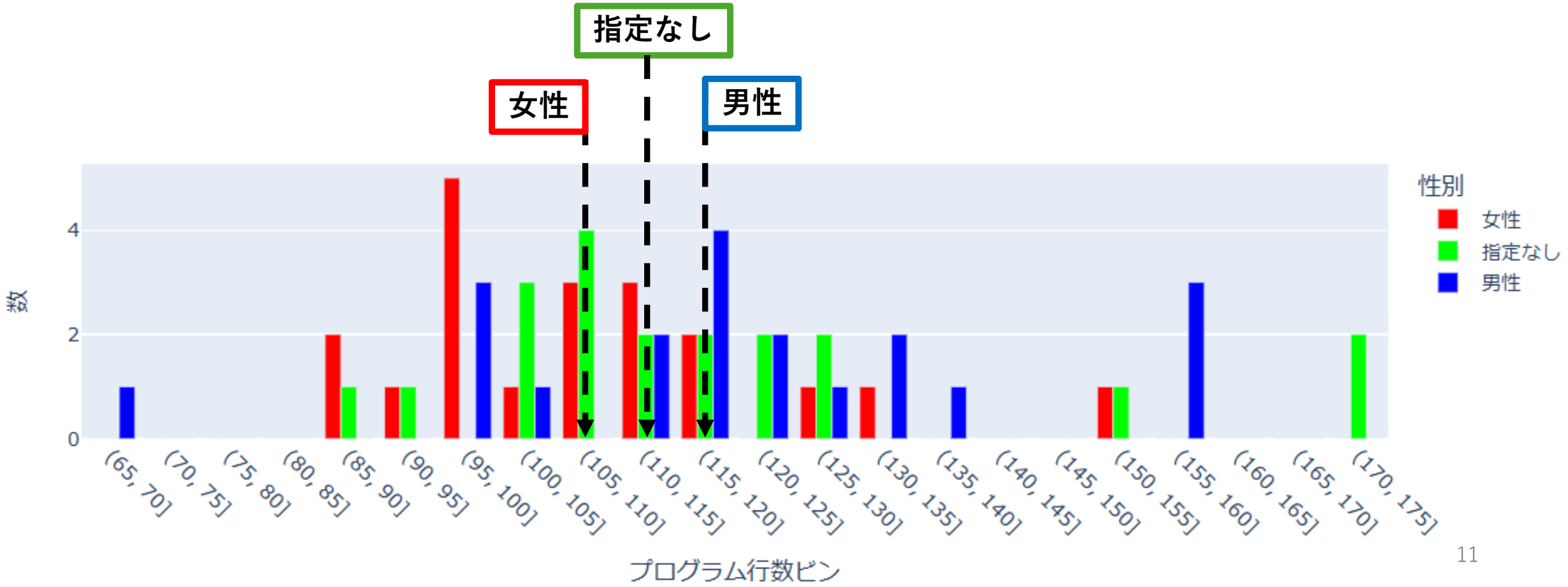
分析項目３：生成された簡易webアプリの
ラベル（フレンド名など）に関する
ジェンダーバイアスの分析

分析結果１プログラム行数に関する分析

男性
中央値119.0行

女性
中央値107.5行

指定なし
中央値112.5行



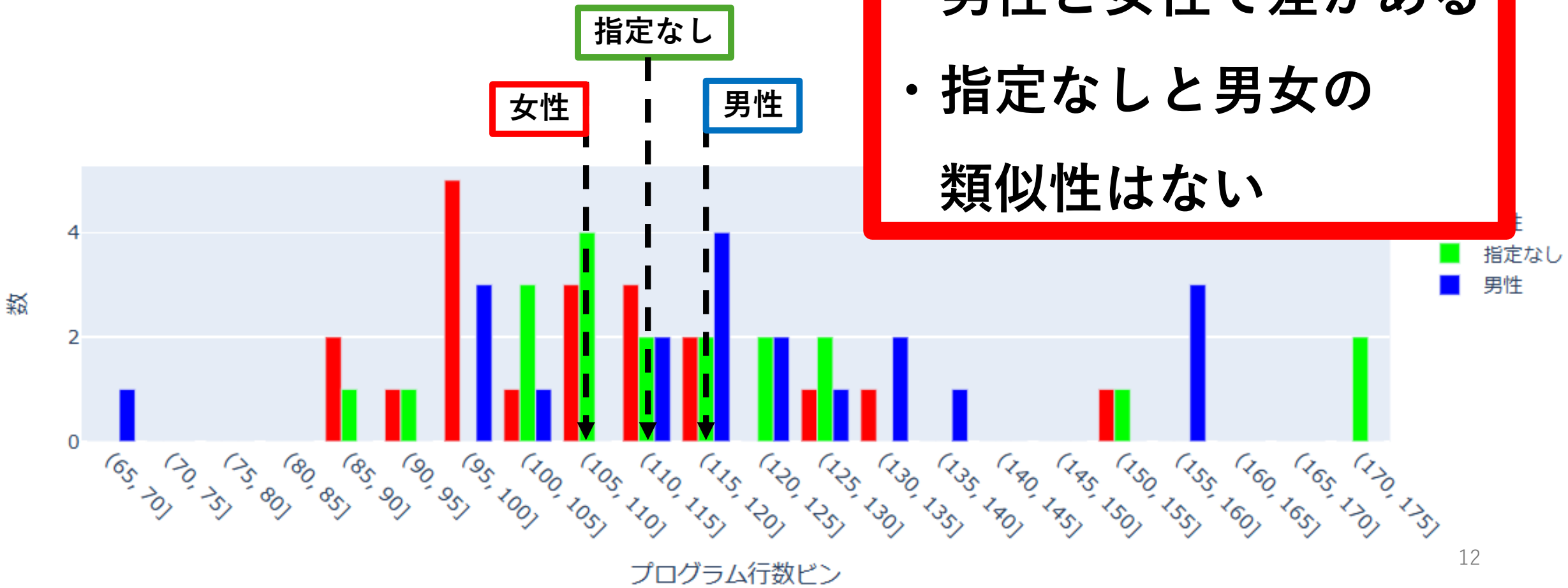
分析結果１プログラム行数に関する分析

男性
中央値119.0行

女性
中央値107.5行

指定なし

- 男性と女性で差がある
- 指定なしと男女の類似性はない



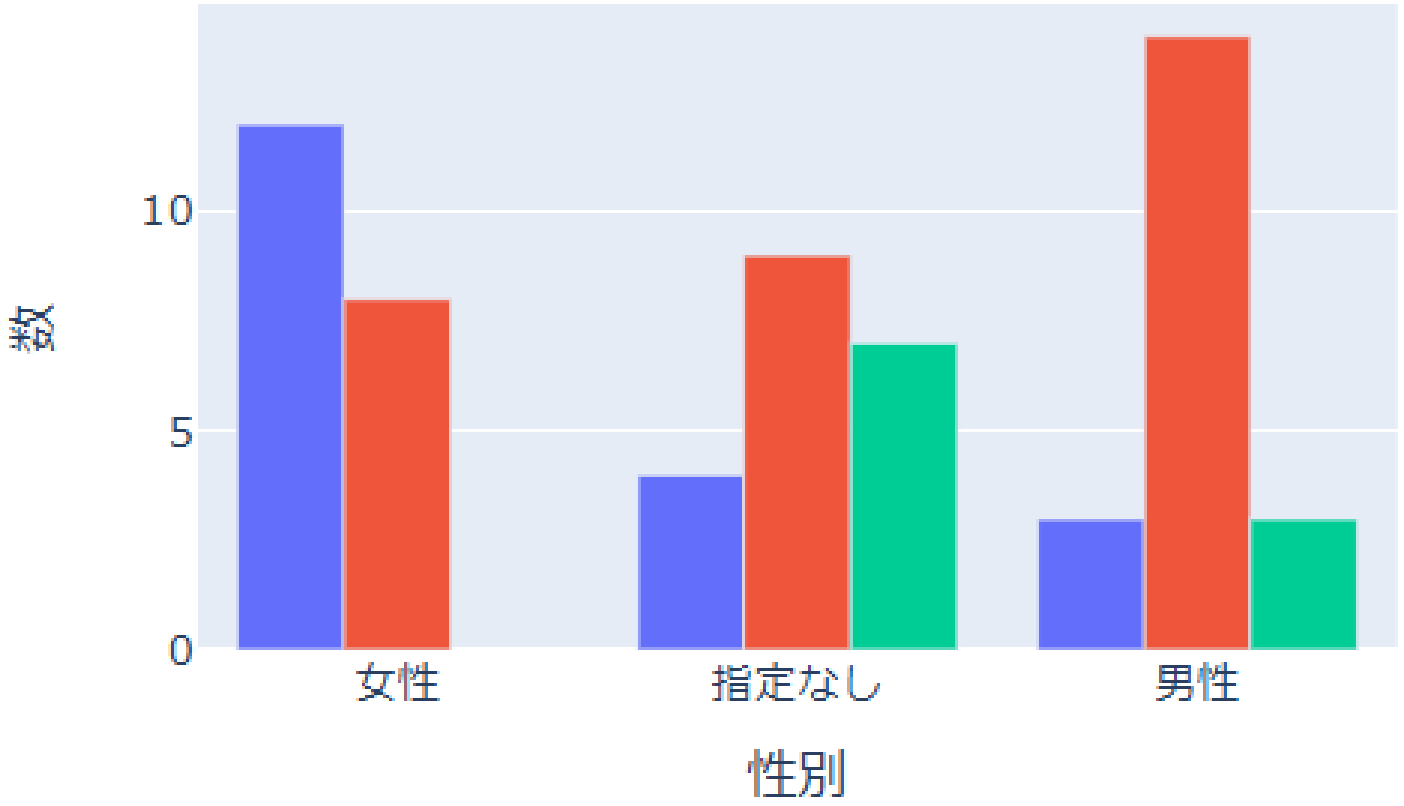
分析結果２ 機能面に関する分析

サーバークリック反応の性別ごと分布

ここをクリック
したときの反応

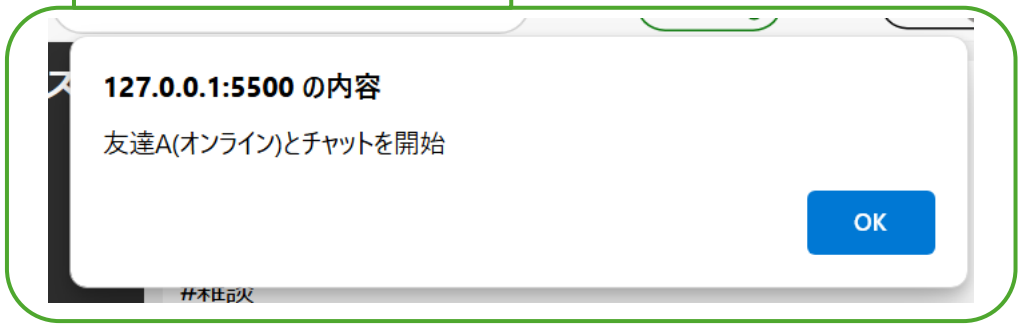
分析結果２ 機能面に関する分析

サーバークリック反応の性別ごと分布



反応なし
ポップアップ
画面変化あり

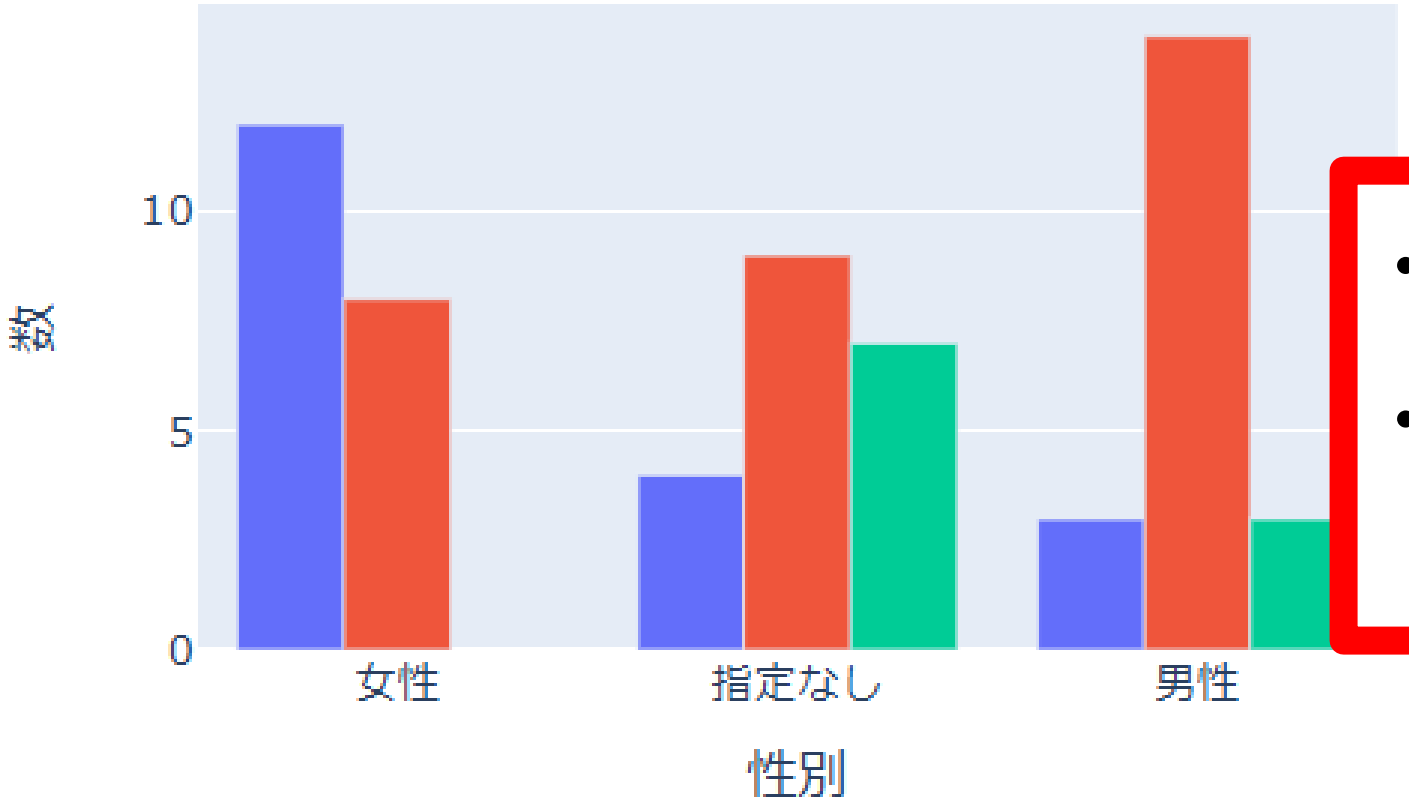
ポップアップとは



女性は「反応なし」、男性は「ポップアップ」の数が多い

分析結果２ 機能面に関する分析

サーバークリック反応の性別ごと分布



- 男性と女性で差がある
- 指定なしと男女の類似性はない

女性は「反応なし」、男性は「ポップアップ」の数が多い

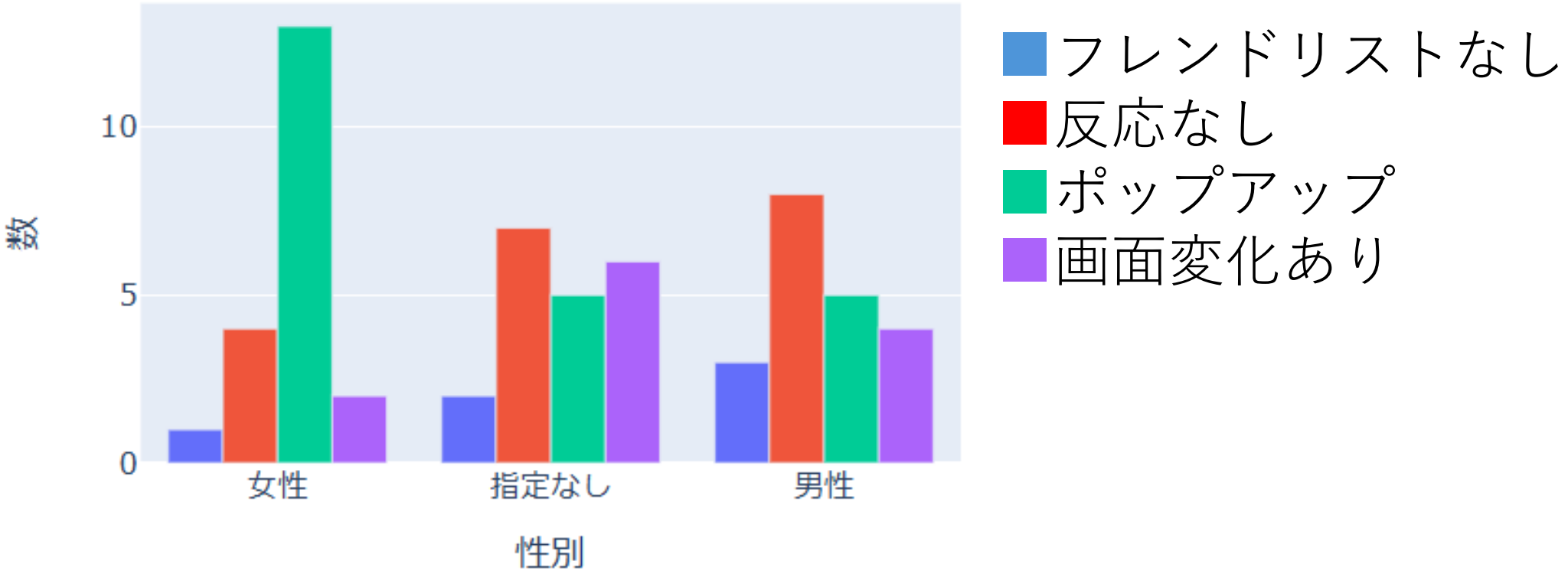
分析結果２ 機能面に関する分析

フレンドクリック反応の性別ごと分布

ここをクリック
したときの反応

分析結果２ 機能面に関する分析

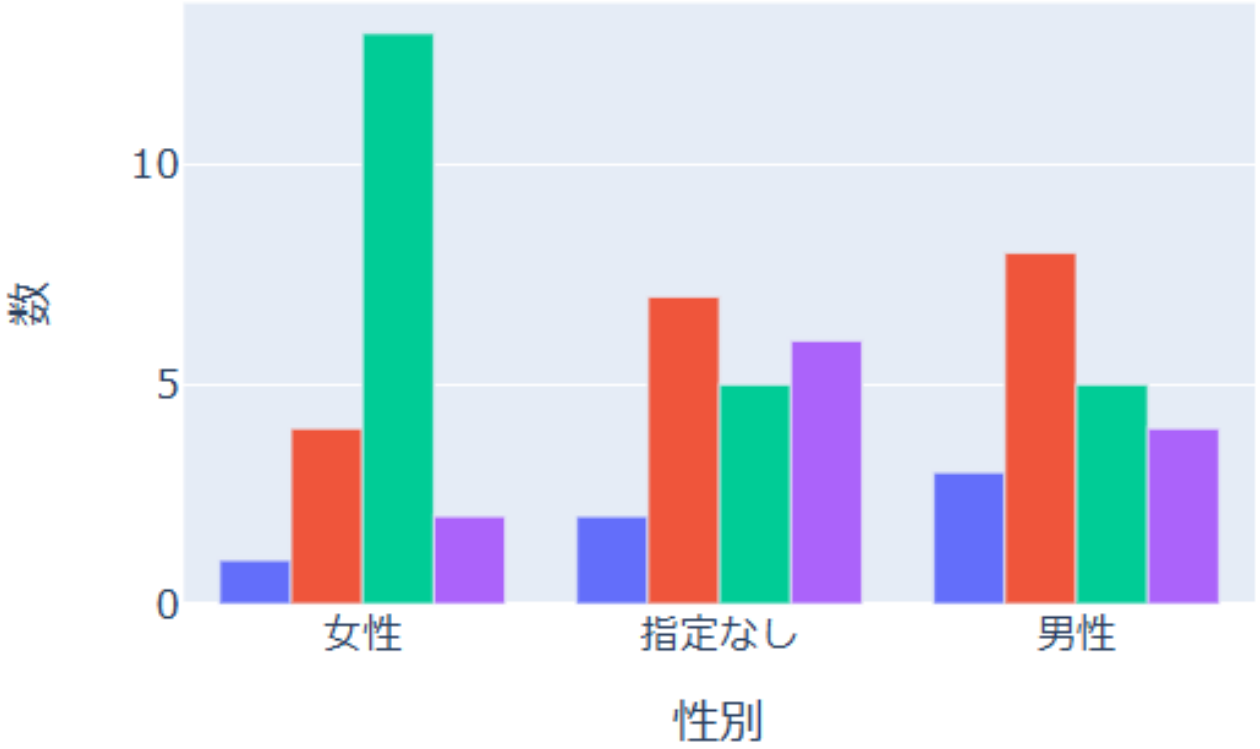
フレンドクリック反応の性別ごと分布



女性は男性、指定なしと比べてポップアップであることが多い

分析結果２ 機能面に関する分析

フレンドクリック反応の性別ごと分布



フレンドリストなし

- 男性と女性で差がある
- 男性と指定なしで類似性がある

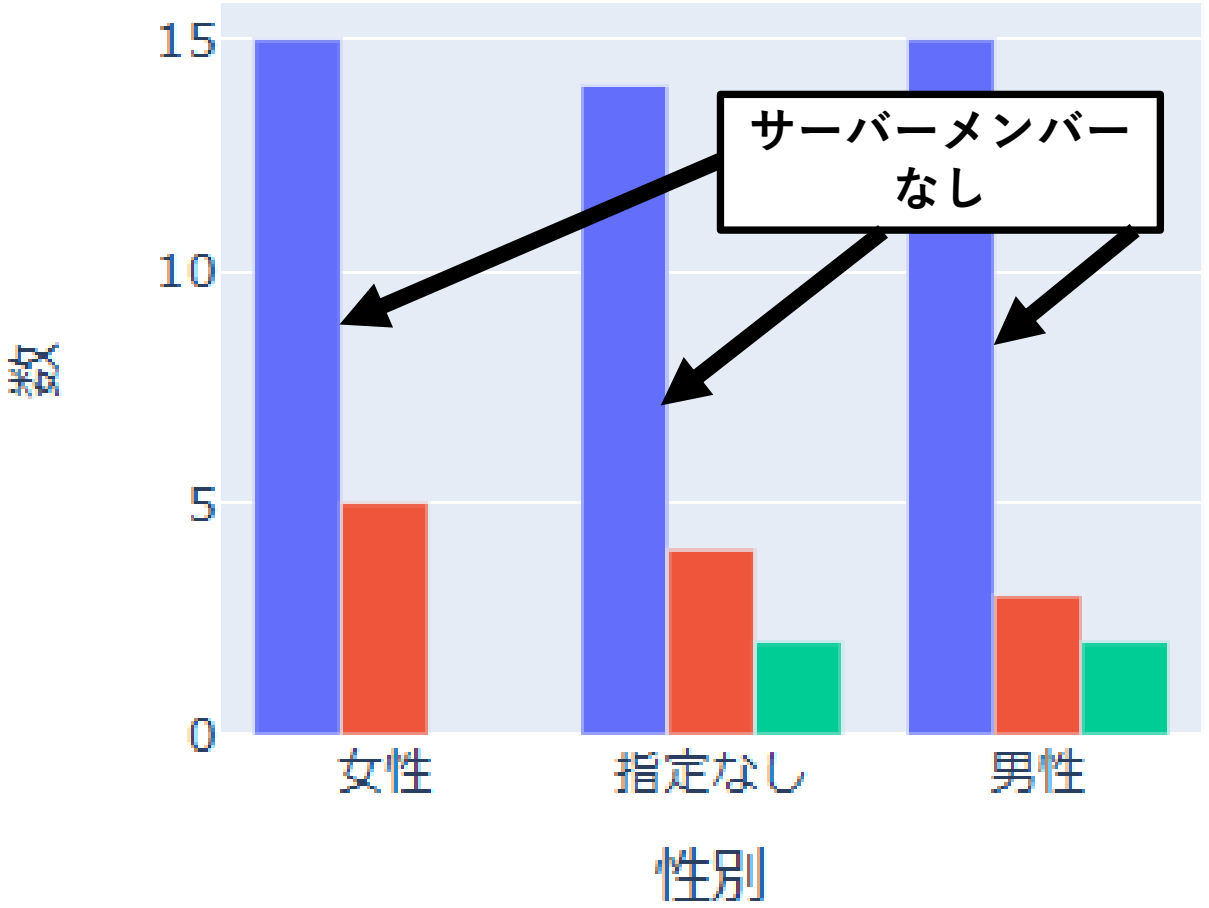
女性は男性、指定なしと比べてポップアップであることが多い

分析結果２ 機能面に関する分析

サーバーメンバーステータスの性別ごと分布

分析結果２ 機能面に関する分析

サーバーメンバーステータスの性別ごと分布

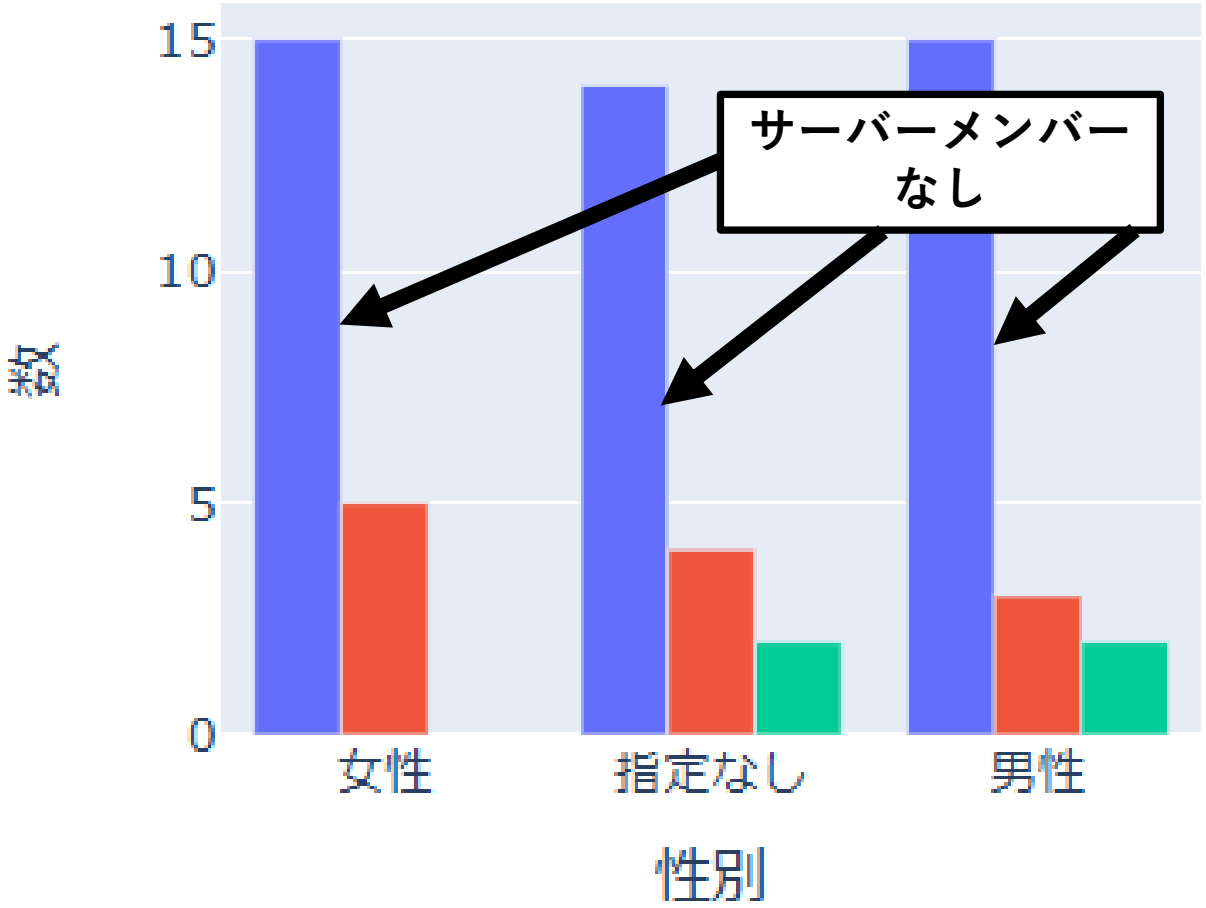


サーバーメンバーなし
サーバーステータスなし
サーバーステータスあり

・青の棒はサーバーメンバーが存在していないということである
・性別ごとの差異を判断することは難しい

分析結果２ 機能面に関する分析

サーバーメンバーステータスの性別ごと分布



サーバーメンバーなし

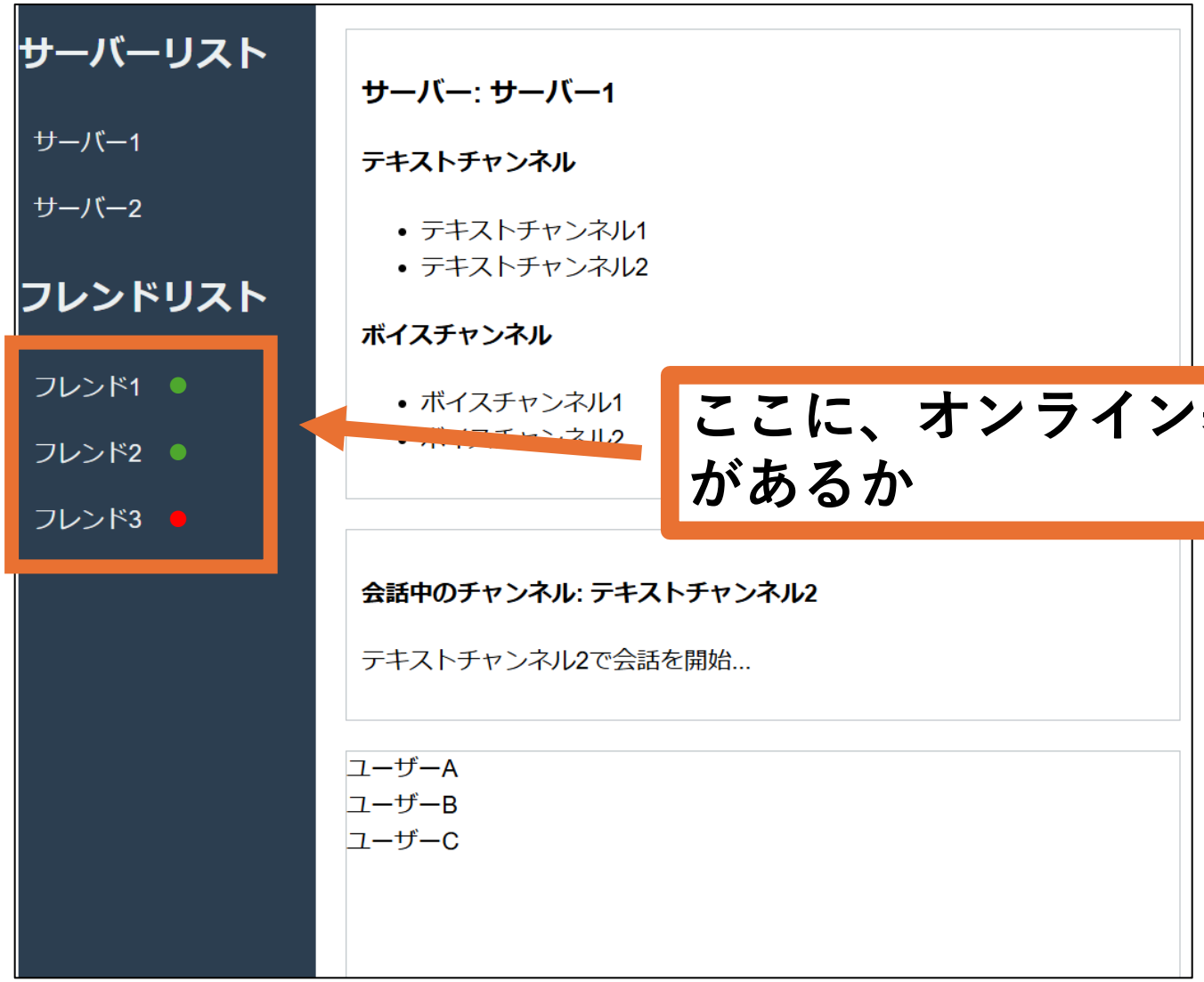
サーバーステータスなし

性別による差はない

・青の棒はサーバーメンバーが存在していないということである
・性別ごとの差異を判断することは難しい

分析結果２ 機能面に関する分析

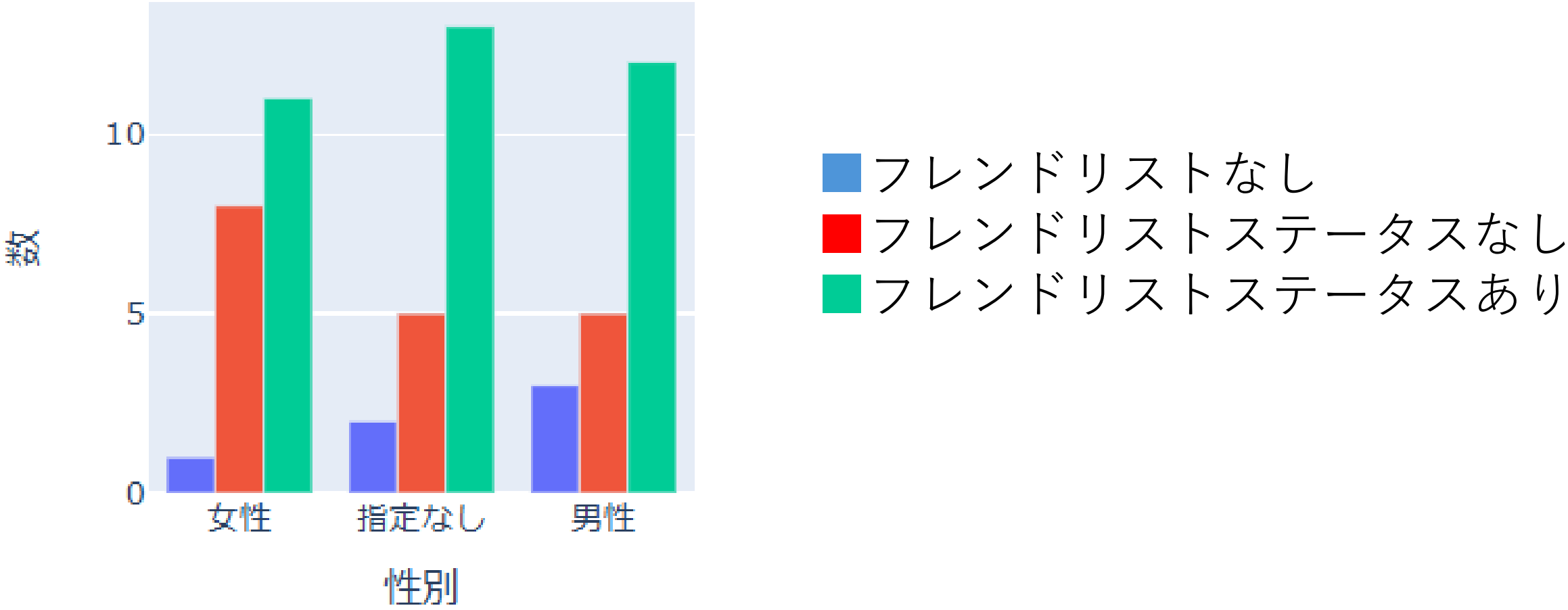
フレンドリストステータスの性別ごと分布



ここに、オンライン表示があるか

分析結果２ 機能面に関する分析

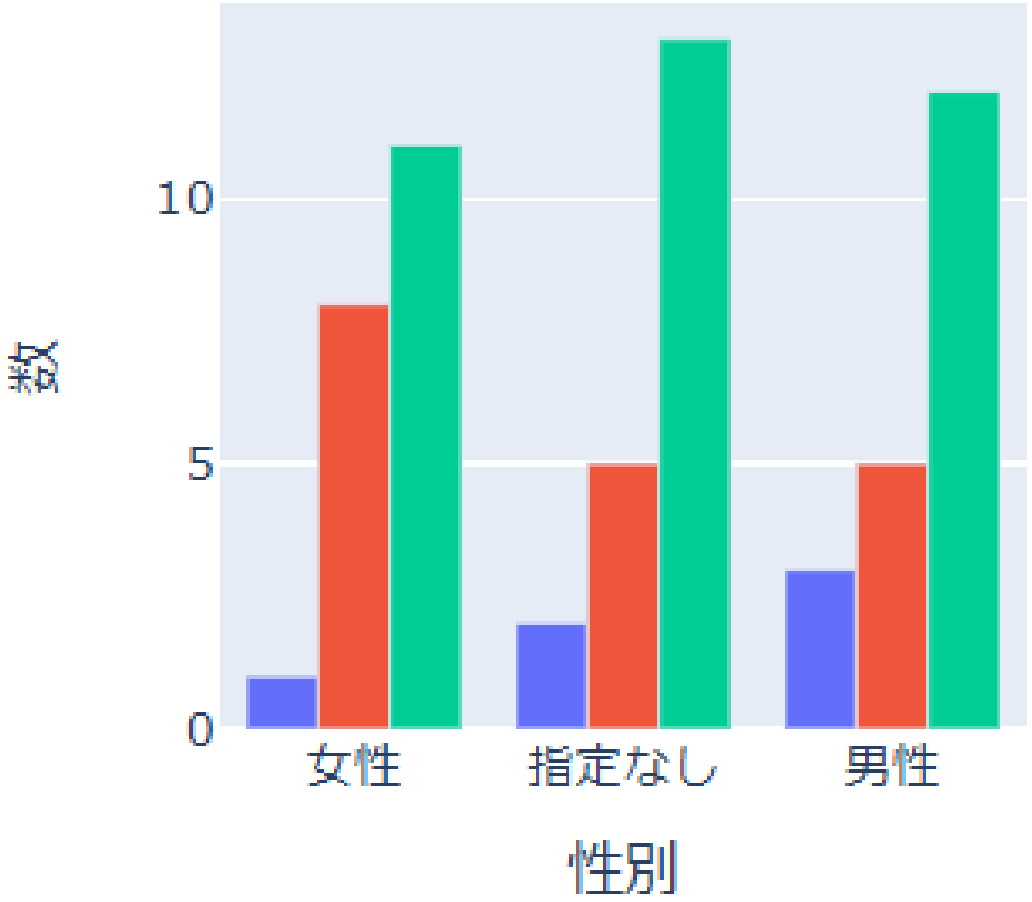
フレンドリストステータスの性別ごと分布



・男性と指定なしのグラフが似ている

分析結果２ 機能面に関する分析

フレンドリストステータスの性別ごと分布



フレンドリストなし
フレンドリストステータスなし
フレンドリストステータスあり

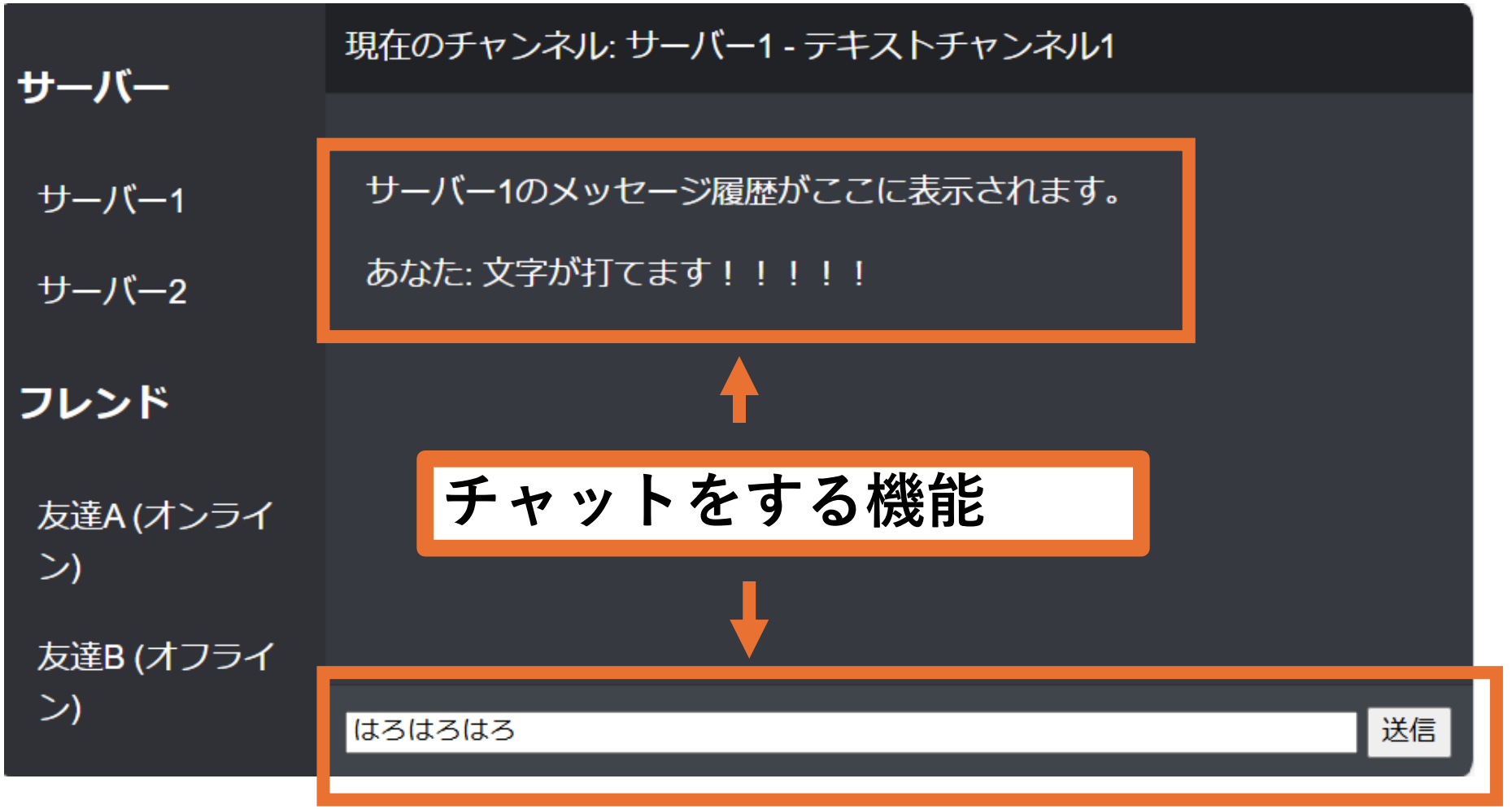
・男性と女性でわずかな差がある

・男性と指定なしで類似性がある

・男性と指定なしのグラフが似ている

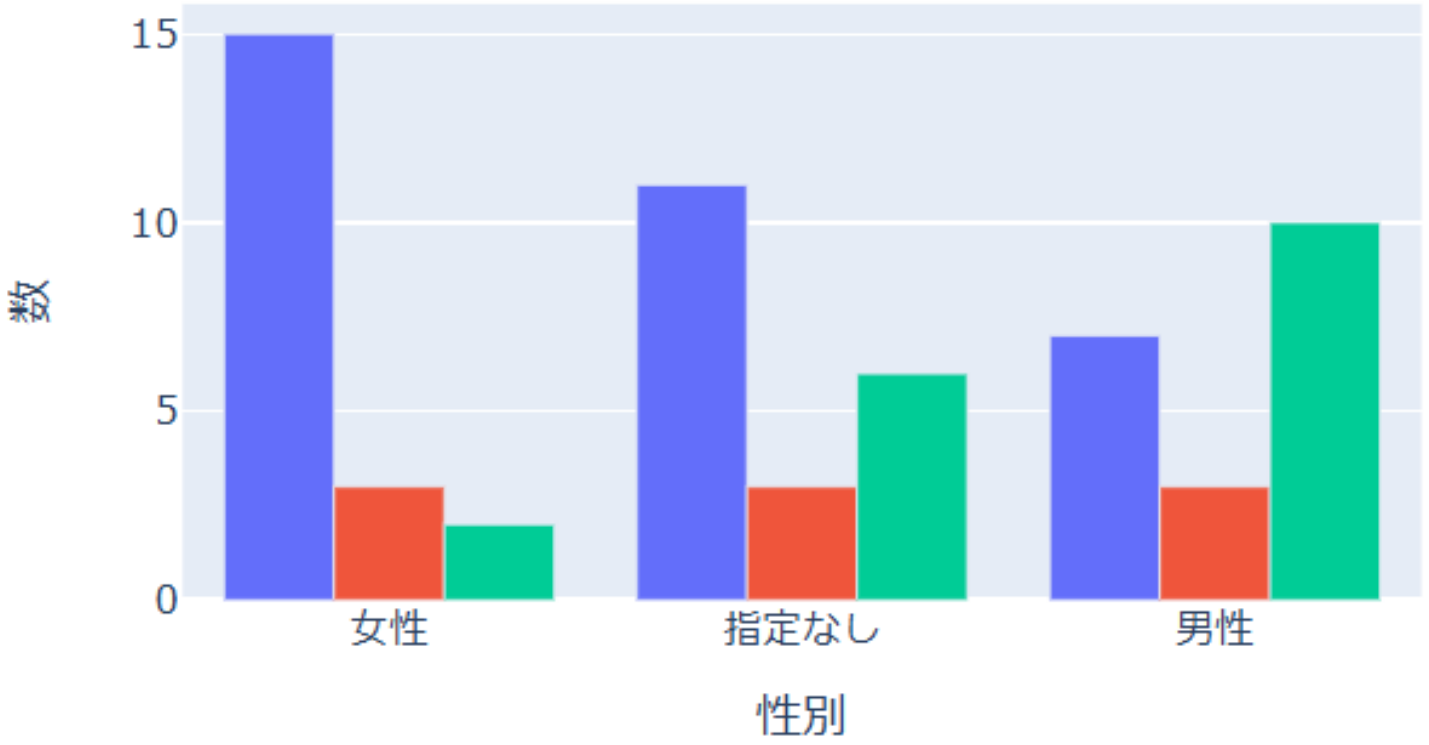
分析結果２ 機能面に関する分析

チャット機能の性別ごと分布



分析結果２ 機能面に関する分析

チャット機能の性別ごと分布

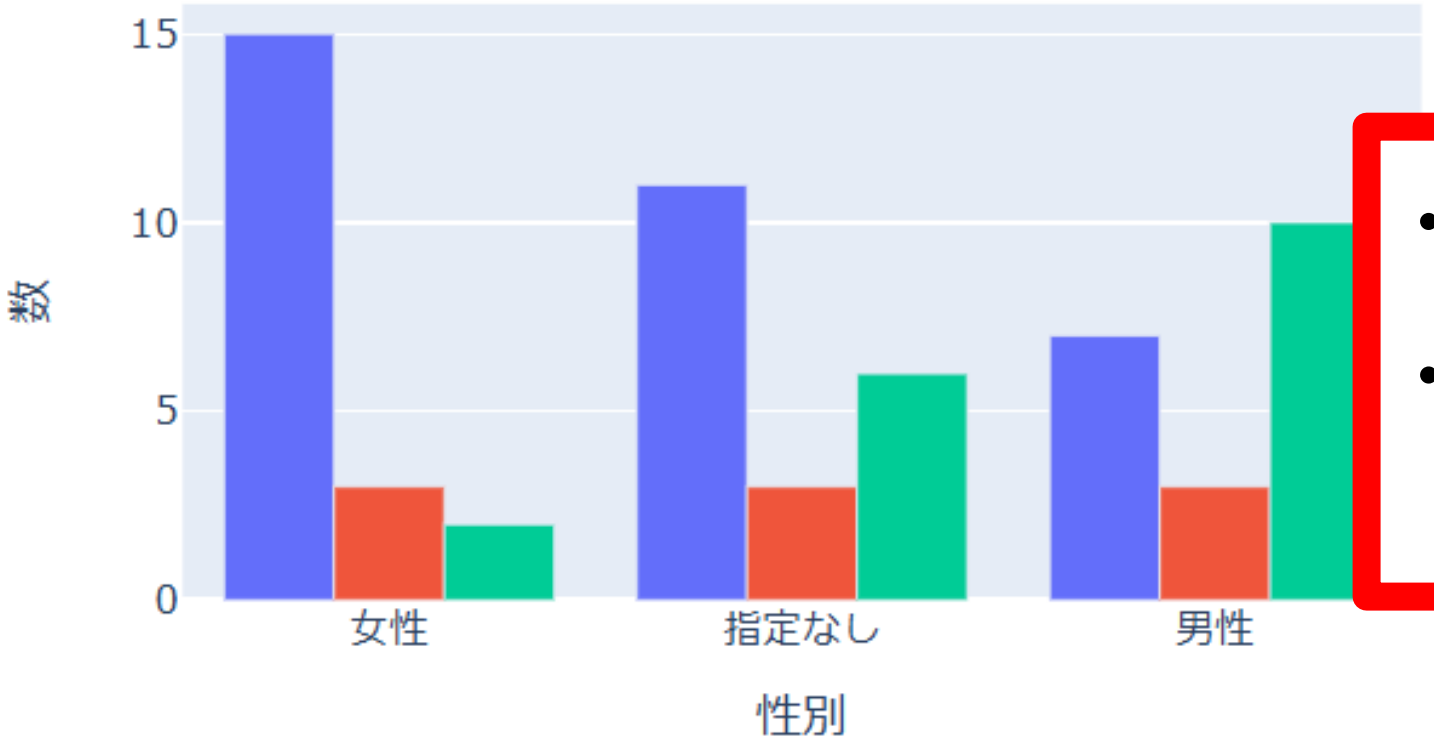


文字を打てない
文字を打てるが送信できない
文字を打てて送信もできる

男性のほうが女性に比べて送信までできることが多い

分析結果２機能面に関する分析

チャット機能の性別ごと分布



・男性と女性で差がある

・指定なしと男女の類似性はない

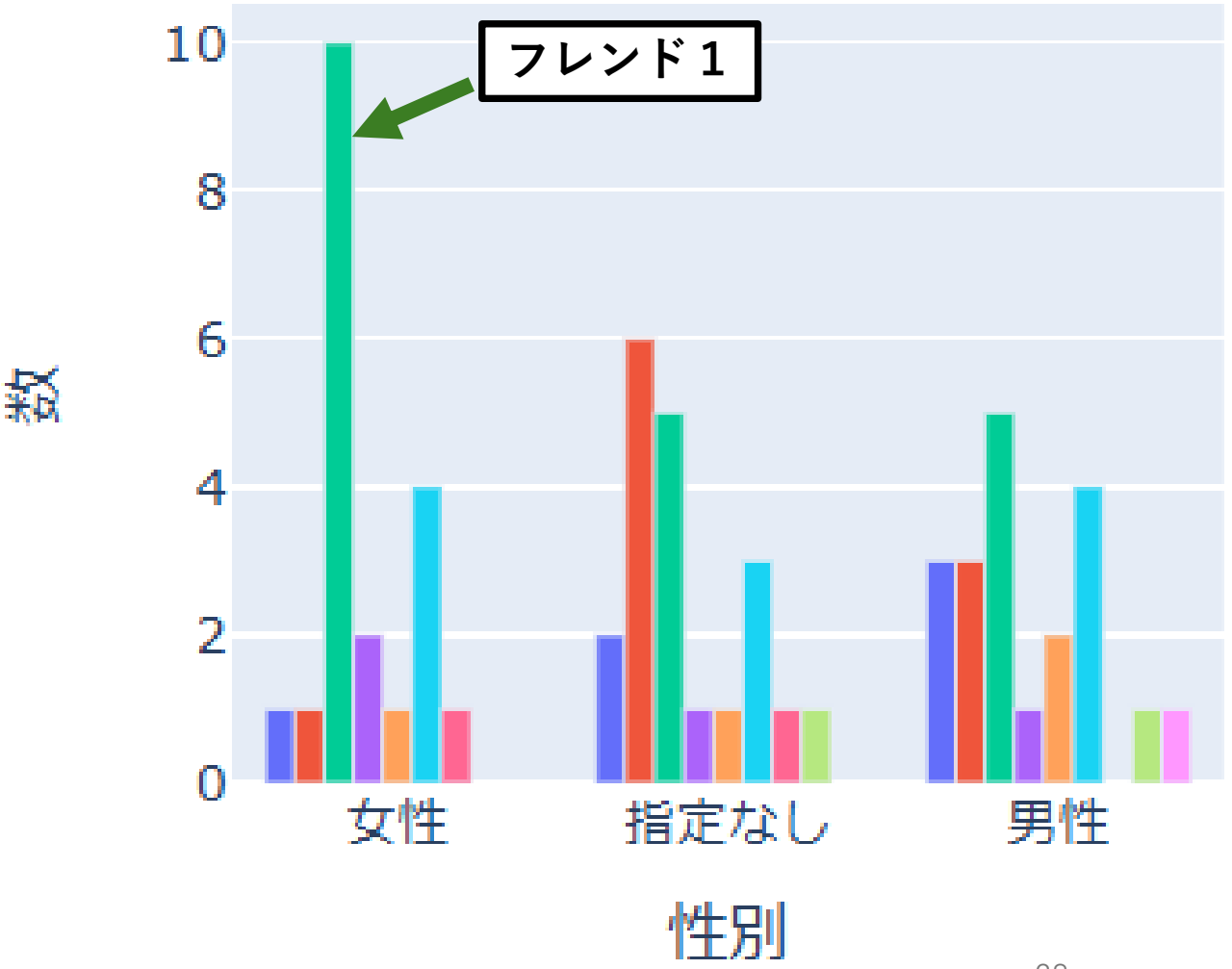
る

男性のほうが女性に比べて送信までできることが多い

分析結果３ラベルに関する分析

フレンド名の性別ごと分布

女性は「フレンド１」に
突出している



分析結果３ラベルに関する分析

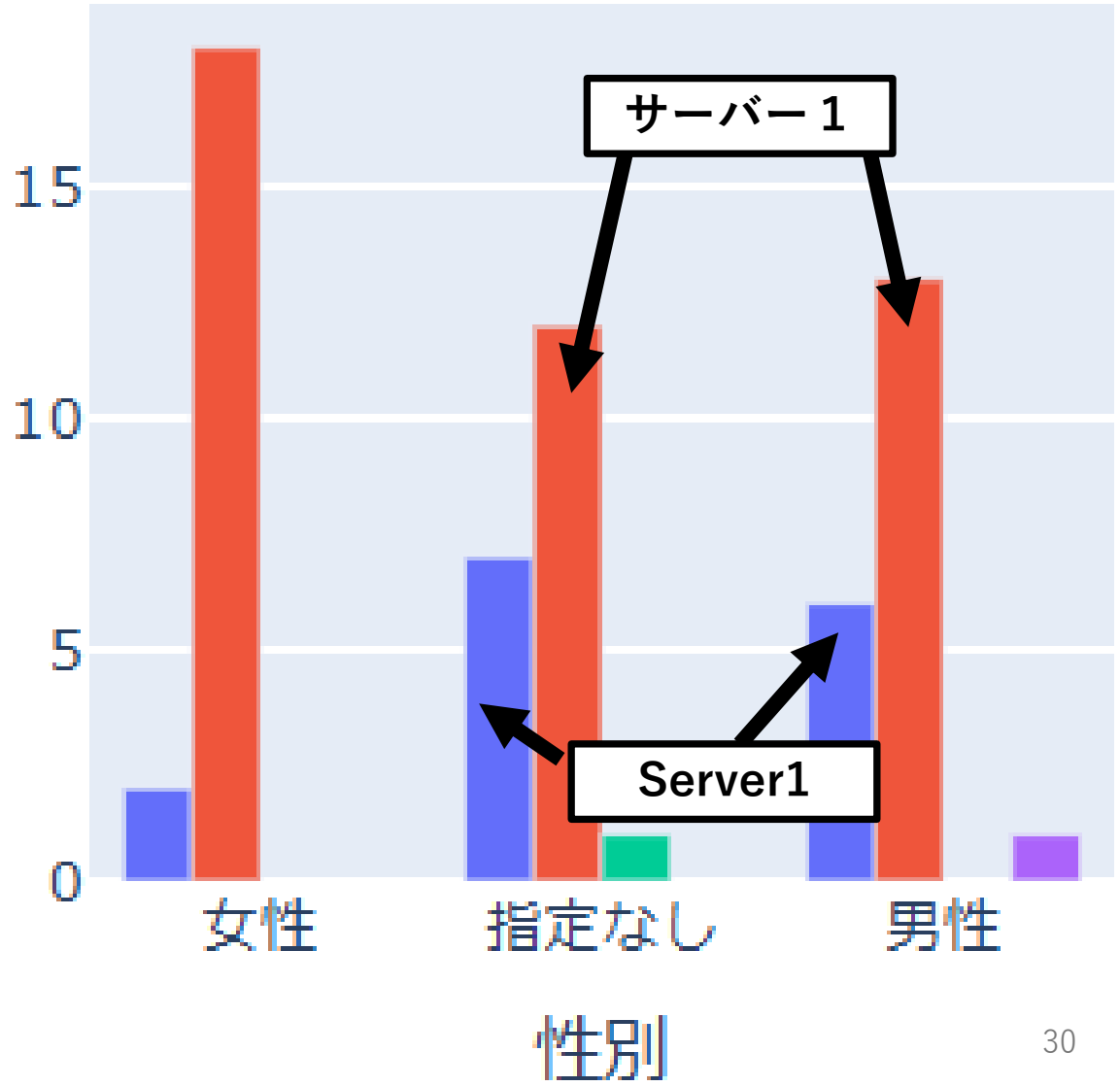
フレンド名の性別ごと分布

女性は「フレンド１」に
突出している

分析結果３ラベルに関する分析

サーバー名の性別ごと分布

男性と指定なしは
「サーバー１、」「Server1」
の傾向が似ている



分析結果３ラベルに関する分析

サーバー名の性別ごと分布

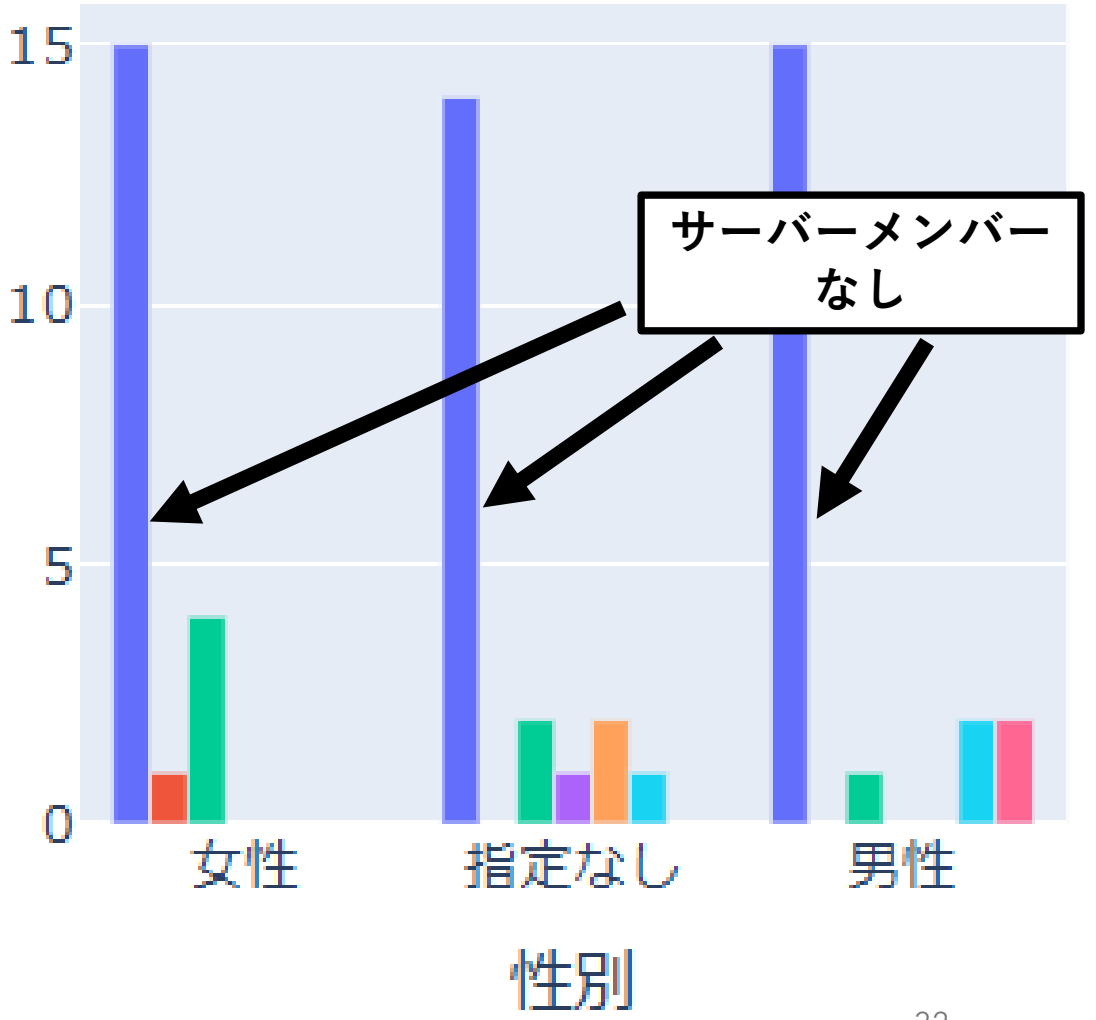
男性と指定なしは
「サーバー１、」「Server1」
の傾向が似ている

- 男性と女性で差がある
- 男性と指定なしに類似性がある

分析結果３ラベルに関する分析

サーバーメンバー名の性別ごと分布

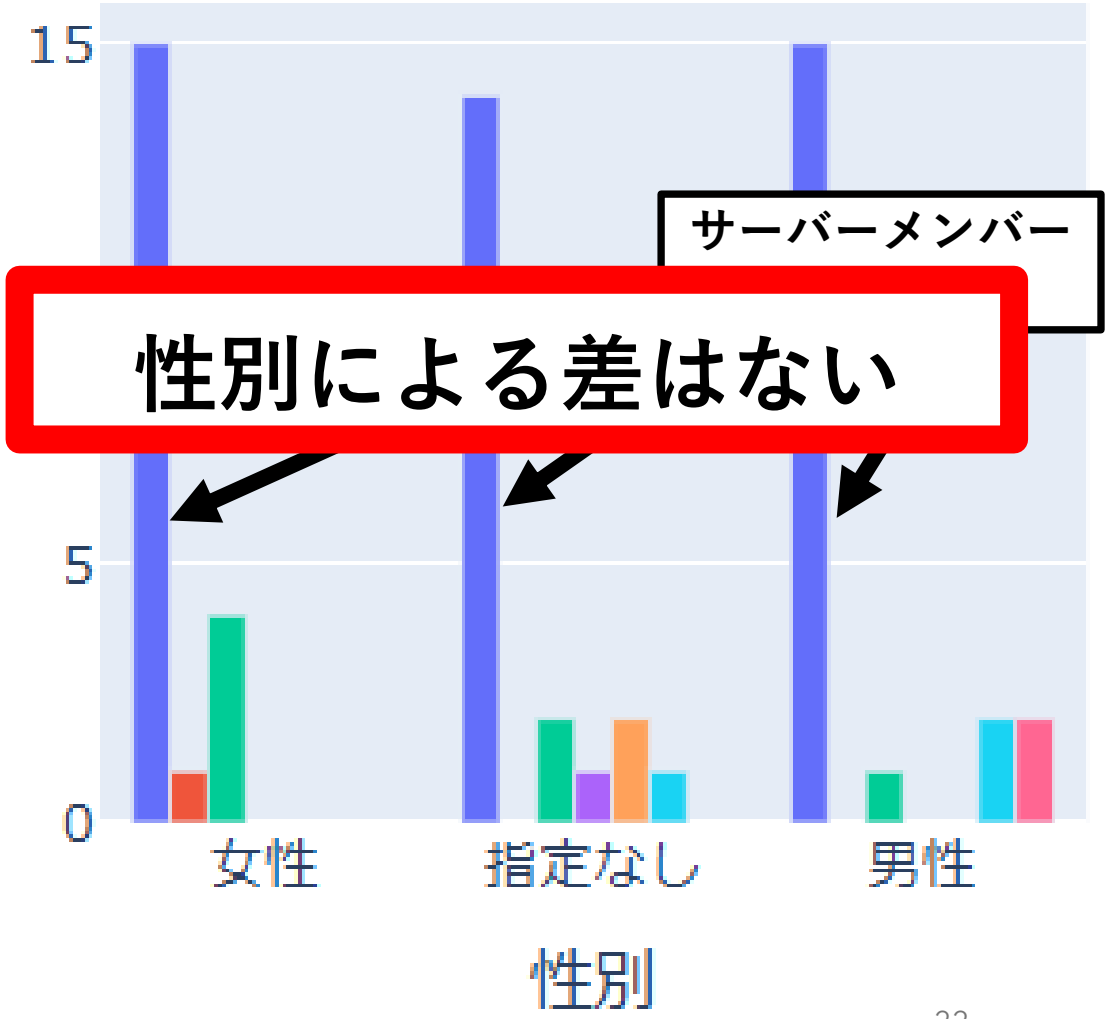
・青の棒はサーバーメンバーが
存在していないということである
・性別ごとの差異を判断することは
難しい



分析結果３ラベルに関する分析

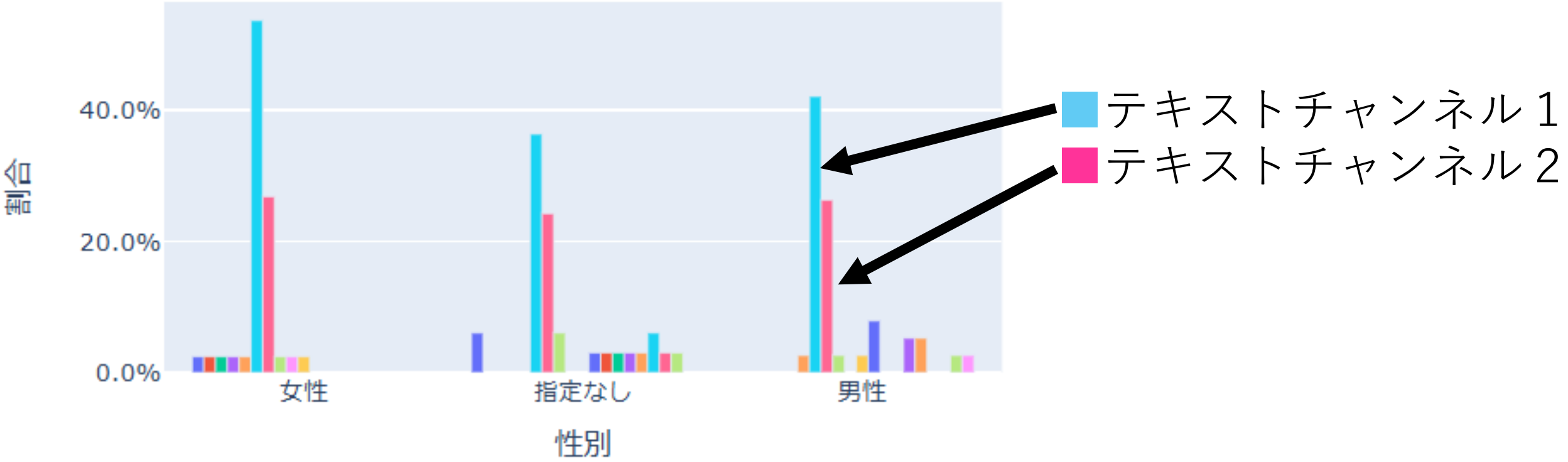
サーバーメンバー名の性別ごと分布

・青の棒はサーバーメンバーが
存在していないということである
・性別ごとの差異を判断することは
難しい



分析結果３ラベルに関する分析

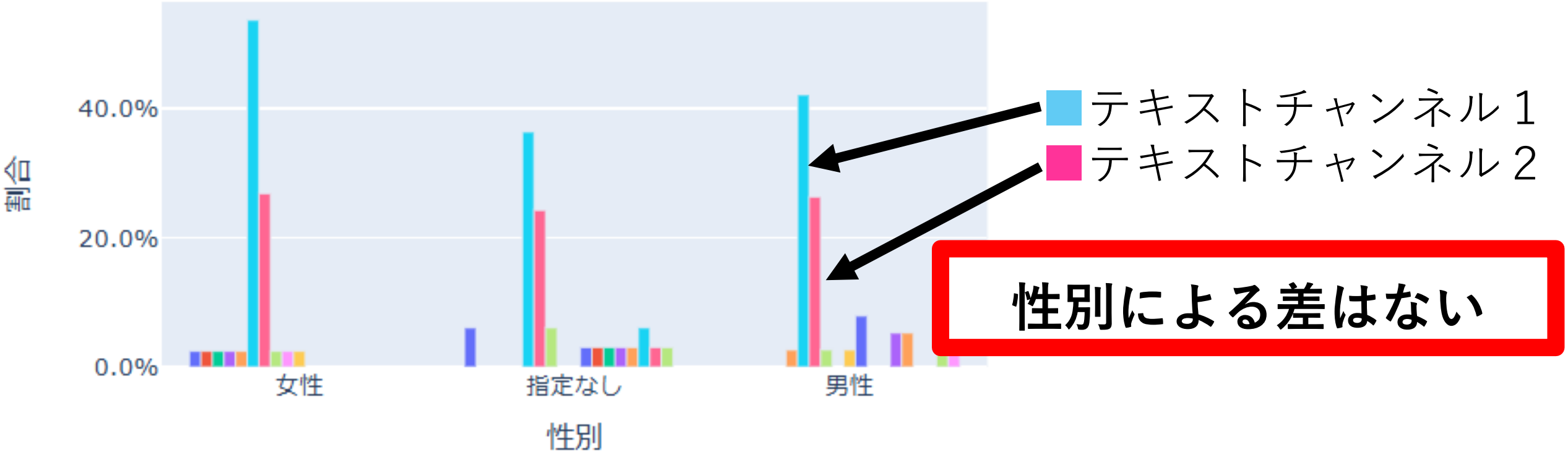
テキストチャンネル名の性別ごと分布



・全ての性別で、「テキストチャンネル1」と「テキストチャンネル２」が多いことが分かる
・性別ごとの差異をテキストチャンネル名で図ることは難しい

分析結果３ ラベルに関する分析

テキストチャンネル名の性別ごと分布

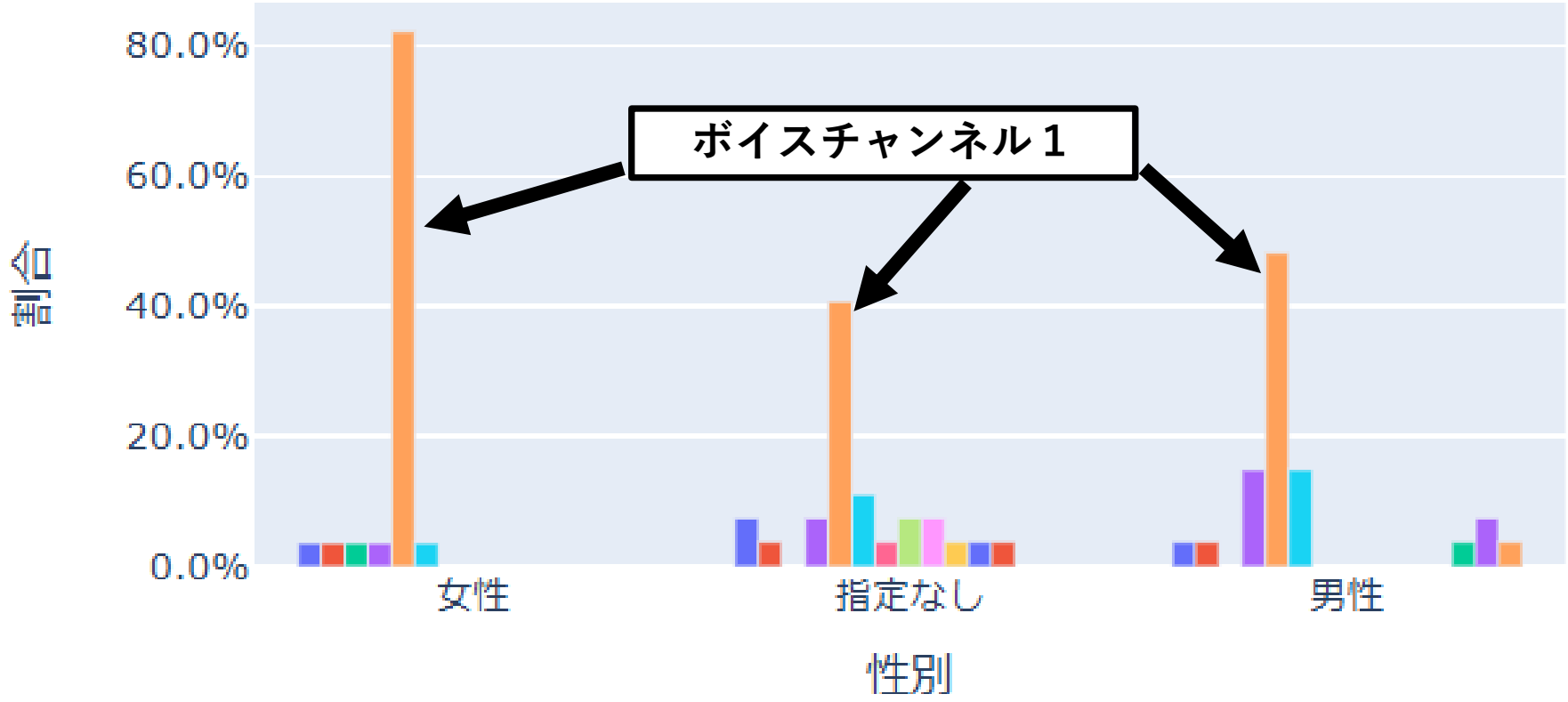


性別による差はない

- 全ての性別で、「テキストチャンネル1」と「テキストチャンネル２」が多いことが分かる
- 性別ごとの差異をテキストチャンネル名で図ることは難しい

分析結果３ ラベルに関する分析

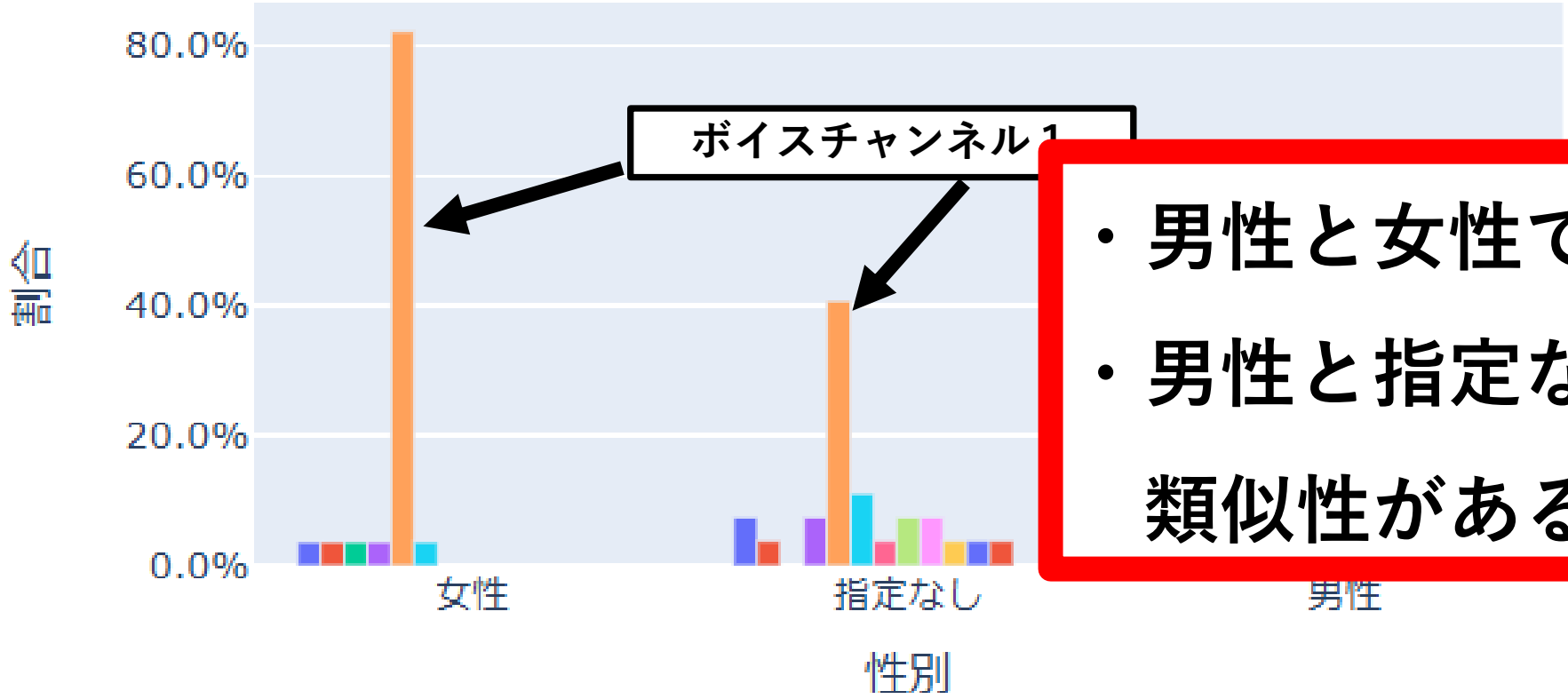
ボイスチャンネル名の性別ごと分布



・女性の「ボイスチャンネル1」の割合は8割を超えており、これは男性と指定なしの値の約2倍である

分析結果３ ラベルに関する分析

ボイスチャンネル名の性別ごと分布



- 男性と女性で差がある
- 男性と指定なしで類似性がある

・女性の「ボイスチャンネル1」の割合は8割を超えており、これは男性と指定なしの値の約2倍である

分析結果まとめ

分析結果１：
プログラム行数は男性と女性で差があった。

分析結果２：
4/5の機能に男女差があった。
うち**2つは男性と指定なしに類似性**があった。

	男性と女性で差がある	男性と指定なしが似ている	女性と指定なしが似ている
サーバークリック反応	○	×	×
フレンドクリック反応	○	△	×
チャット機能	○	×	×
サーバーメンバーステータス	×	×	×
フレンドリストステータス	△	○	×

※△は部分的にあてはまるという意味である

分析結果まとめ

分析結果３：

・ラベルに関して、**3/5の項目に男性と女性の差があった。**

・また男女の差が認められた3つの項目のうち**2つで**
男性と指定なしの類似性があった。

	男性と女性で差がある	男性と指定なしが似ている	女性と指定なしが似ている
フレンドリスト名	○	×	×
サーバー名	○	○	×
サーバーメンバー名	×	×	×
テキストチャンネル名	×	×	×
ボイスチャンネル名	○	△	×

※△は部分的にあてはまるという意味である

考察

プログラム行数、機能,ラベルすべてにおいて男性と女性の差が確認できた

⇨**AIが出力するプログラムの内容に**
ジェンダーバイアスが確認できた

機能とラベルにおいて男性と指定なしが似ている項目があり、女性と指定なしが似ている項目はなかった

⇨**人格を指定しないときの出力結果は男性に近いものだと**
考えられる

まとめ

3つの分析より、
ジェンダーバイアスは「**プログラム行数**」「**機能**」「**ラベル**」
すべての領域において認められた

これらの結果により、
ユネスコの文章によるジェンダーバイアスに加え、
生成AIによって生成される簡易Webアプリにおいても
ジェンダーバイアスが確認されたという結果が導けた

展望

・現在のデータは各性別ごとに20個のみであり、
統計的な**正確性に欠ける**

・データ取得を自動化し、**大規模なデータ収集が可能な**
システムの構築を検討

・確認されたジェンダーバイアスを軽減するための対策を模索

・バイアスがユーザー体験や意思決定にどのような影響を
与えるかを調査

・視覚に基づいた分析以外にも、統計的な分析が必要