JMRAイノベーションキャスト【B3】

抽出率を補正する層別リサンプリング
－ 割り付け標本の精度向上提案 －

リサーチ・イノベーション委員会

相澤 健

プロフィール

講演者：
相澤　健　　株式会社サイズ　営業本部

ナビゲーター：
朝野 熙彦　　都立大学・専修大学元教授

著書は「ディープラーニングwith Python」
「最新マーケティング・サイエンスの基礎」など

本日の論点

- 割り付け標本から生じる歪みを層別リサンプリングで補正する方法を提案する
- 確率分布の仮定を入れないノンパラメトリックなデータ生成法である
- クロス集計補正の意義と今後の拡張可能性を論じる

提 言

- 調査の目的と予算等の制約に応じて、各年代セルを均等に割り付けた調査が行われることがある
- 割り付け調査で得られたデータ行列を元に、母集団の構成比により近い単純集計とクロス集計を得る新しい方法を提案する
- **層別リサンプリング**の方法及び検証結果を示す

割り付け調査（サンプルサイズ）							
男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
100	100	100	100	100	100	100	100

割り付け調査から
母集団の構成比に近い
結果を得られるか？

母集団の構成比（%）							
男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
5.1	11.7	16.6	17.3	5.2	12.2	15.3	16.5

従来の方法：性年代別にウェイトづけた補正

Q.好きなテレビ番組のジャンルは？

	割り付け調査の結果		標本調査（正解データ）		割り付け調査結果の補正	
	① 回収サンプルサイズ	② 「ワイドショー」 出現数	③ 回収サンプルサイズ	「ワイドショー」 出現数	ウェイト値 （③/③合計*①合計/①）	ウェイト後の出現数 （②*ウェイト値）
男性12-19才	100	4	260	11	0.41	1.6
男性20-34才	100	5	591	38	0.93	4.7
男性35-49才	100	15	839	113	1.33	19.9
男性50-69才	100	23	877	192	1.39	31.9
女性12-19才	100	4	265	8	0.42	1.7
女性20-34才	100	13	619	82	0.98	12.7
女性35-49才	100	23	777	196	1.23	28.2
女性50-69才	100	41	834	327	1.32	54.0
合計	800	128	5,062	967		154.8
割合（%）	回収データ	16.0	正解データ	19.1	補正データ	19.3

単純集計はウェイトによって補正できた。しかし、データ行列そのものが補正されたわけではない。

層別リサンプリングの方法

- 得られた割り付け調査のデータから、目標とする構成比に合わせて復元抽出を繰り返す
- 層別リサンプリングでは、たとえば男性12-19才は割り付け100サンプルからランダムに１人ずつ取り出し、目標とするサイズの260になるまで繰り返す。他の年代セルについても同様。
- ローデータから１人のデータを抜き出し、それを複数重ねることで新たなローデータを生成

割り付け調査（サンプルサイズ）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
100	100	100	100	100	100	100	100

リサンプリング

目標とする構成比

目標とする構成比（母集団の構成比）（%）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
5.1	11.7	16.6	17.3	5.2	12.2	15.3	16.5

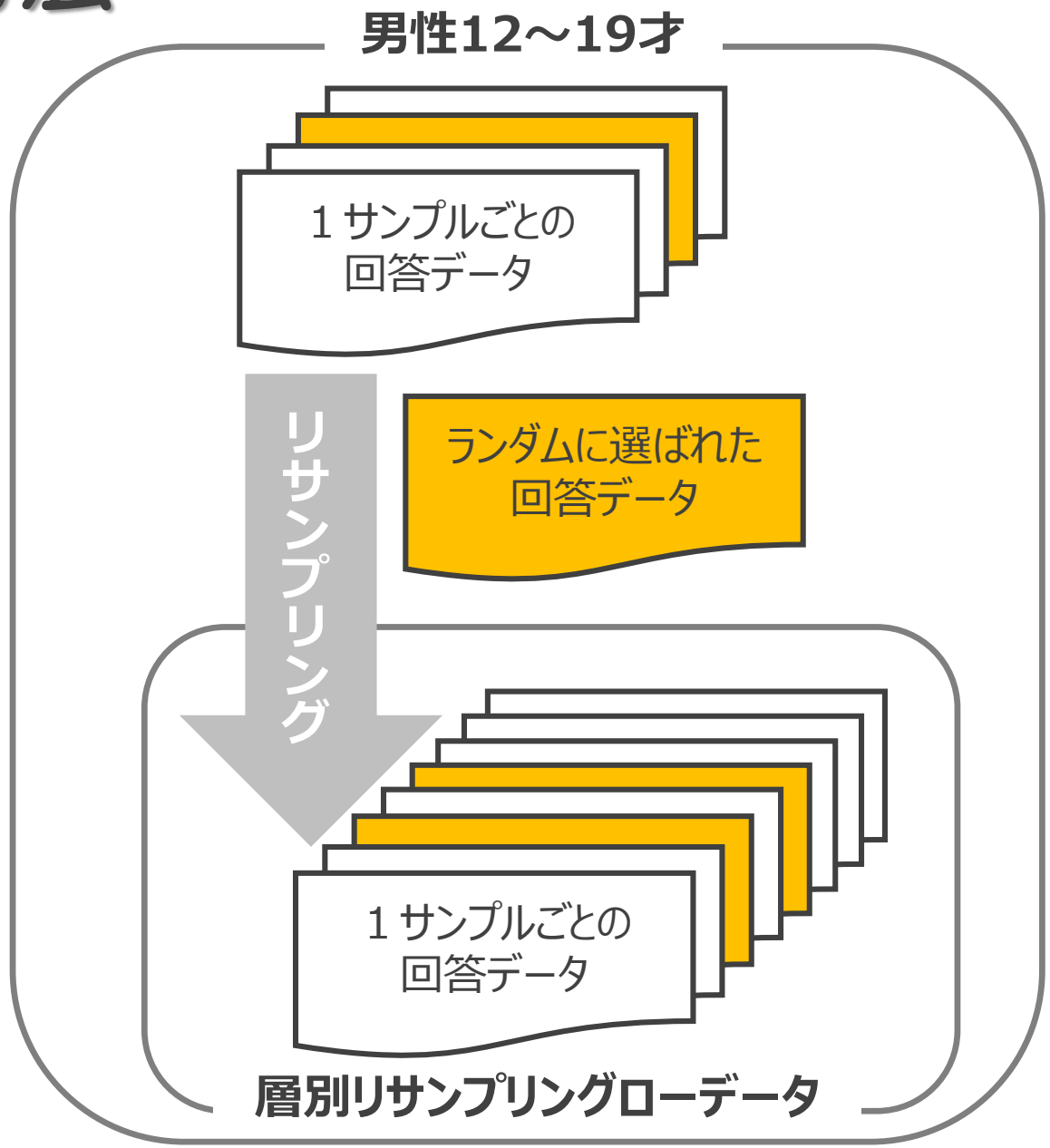
再現するサンプルサイズ

層別リサンプリング（サンプルサイズ）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
260	591	839	877	265	619	777	834

層別リサンプリングの方法

- 得られた割り付け調査のデータから、目標とする構成比に合わせて復元抽出を繰り返す
- 層別リサンプリングでは、たとえば男性12-19才は割り付け100サンプルからランダムに１人ずつ取り出し、目標とするサイズの260になるまで繰り返す。他の年代セルについても同様。
- ローデータから１人のデータを抜き出し、それを複数重ねることで新たなローデータを生成



層別リサンプリングの検証方法

- 検証実験のために、仮想的に大型の調査データが母集団のデータであるとみなし、5062サンプルの回答データが母集団の回答分布であると「想定」
- その大規模な標本調査＝マスターデータから各年代セル100で疑似的に割り付け調査データを抽出
- 疑似的な割り付け調査データから層別リサンプリングでデータ行列を生成
- 疑似的な割り付け調査とそれを元にした層別リサンプリングでどちらの集計値がマスターデータに近いかを評価

マスターデータ（母集団と想定）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
260	591	839	877	265	619	777	834

サンプリング（非復元抽出）

疑似的な割り付け調査（サンプルサイズ）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
100	100	100	100	100	100	100	100

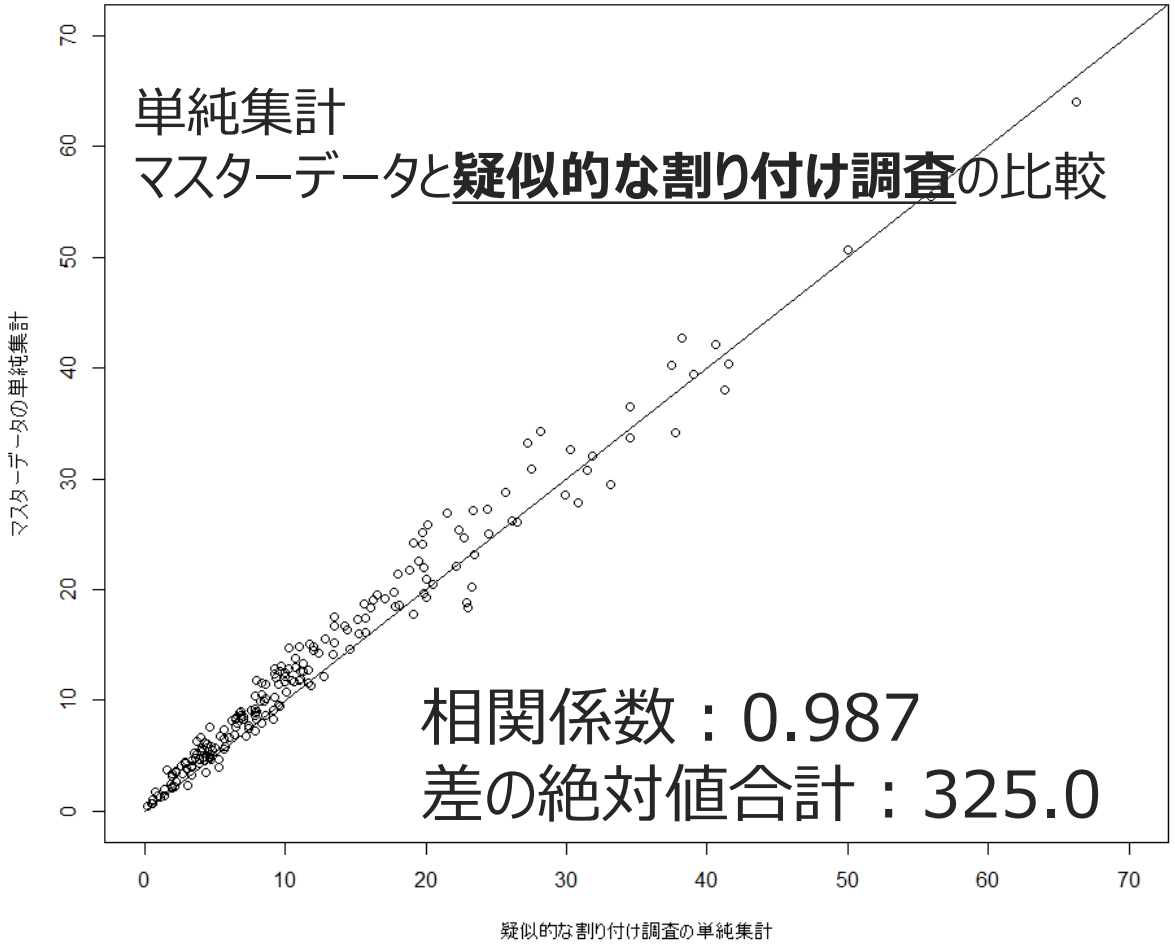
リサンプリング（復元抽出）

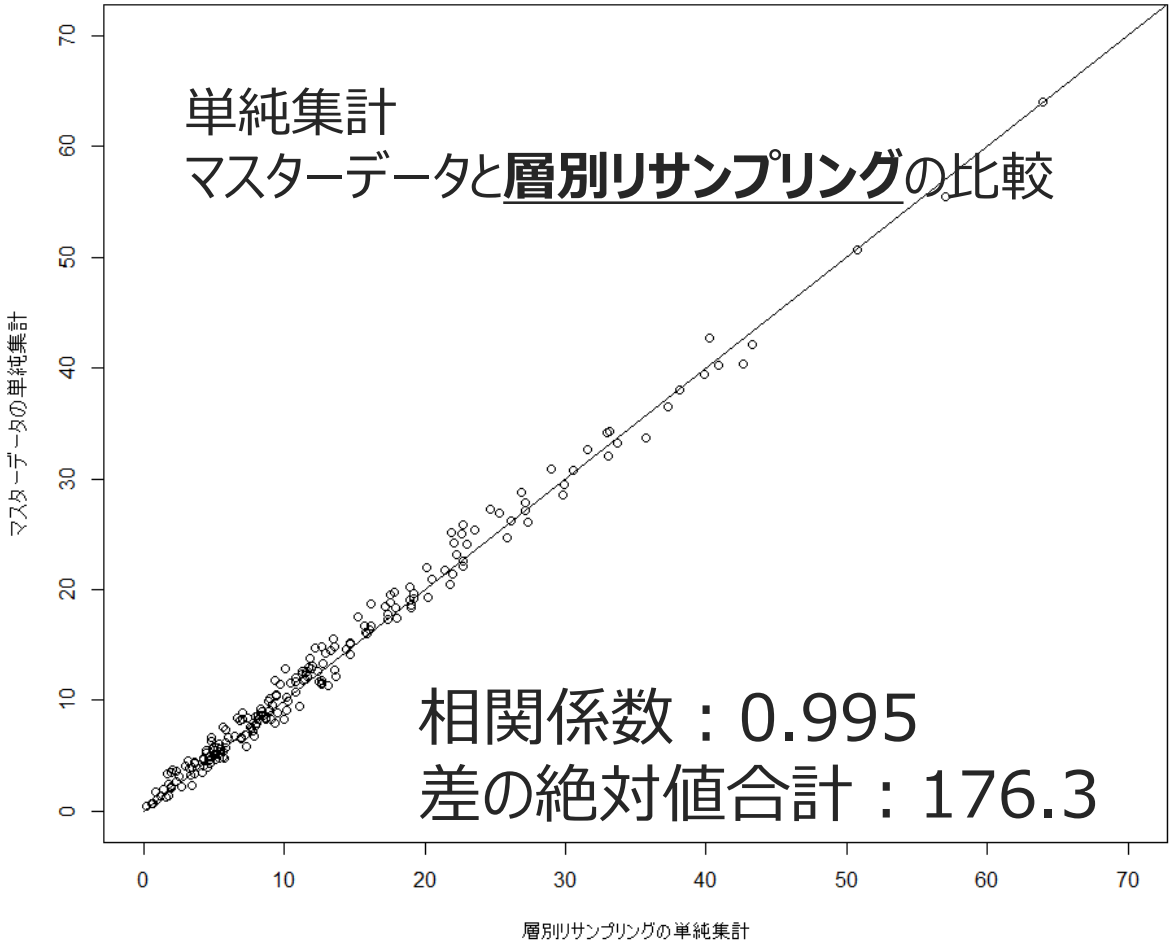
層別リサンプリング（サンプルサイズ）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
260	591	839	877	265	619	777	834

単純集計の比較

- 疑似的な割り付け調査と層別リサンプリングの単純集計結果を散布図で比較（各1回サンプリング）
- 202項目をプロット





クロス集計の比較

- 疑似的な割り付け調査と層別リサンプリングのクロス集計結果を比較
  （各1回サンプリング）

Q１．国産普通乗用車に関心がある？
Q２．好きなテレビ番組ジャンルはニュース番組？

マスターデータ		Q1	
		YES	NO
Q２	YES	10.4	23.8
	NO	10.9	54.8

%

割り付け調査		Q1	
		YES	NO
Q２	YES	7.6	20.5
	NO	10.4	61.5

%

割り付け調査 ⇒

マスターデータとの差（マスターデータ - 割り付け）		Q1	
		YES	NO
Q２	YES	+2.8	+3.3
	NO	+0.6	-6.7

差の絶対値計：13.4

層別リサンプリング		Q1	
		YES	NO
Q２	YES	9.9	23.2
	NO	12.0	54.8

%

層別リサンプリング ⇒

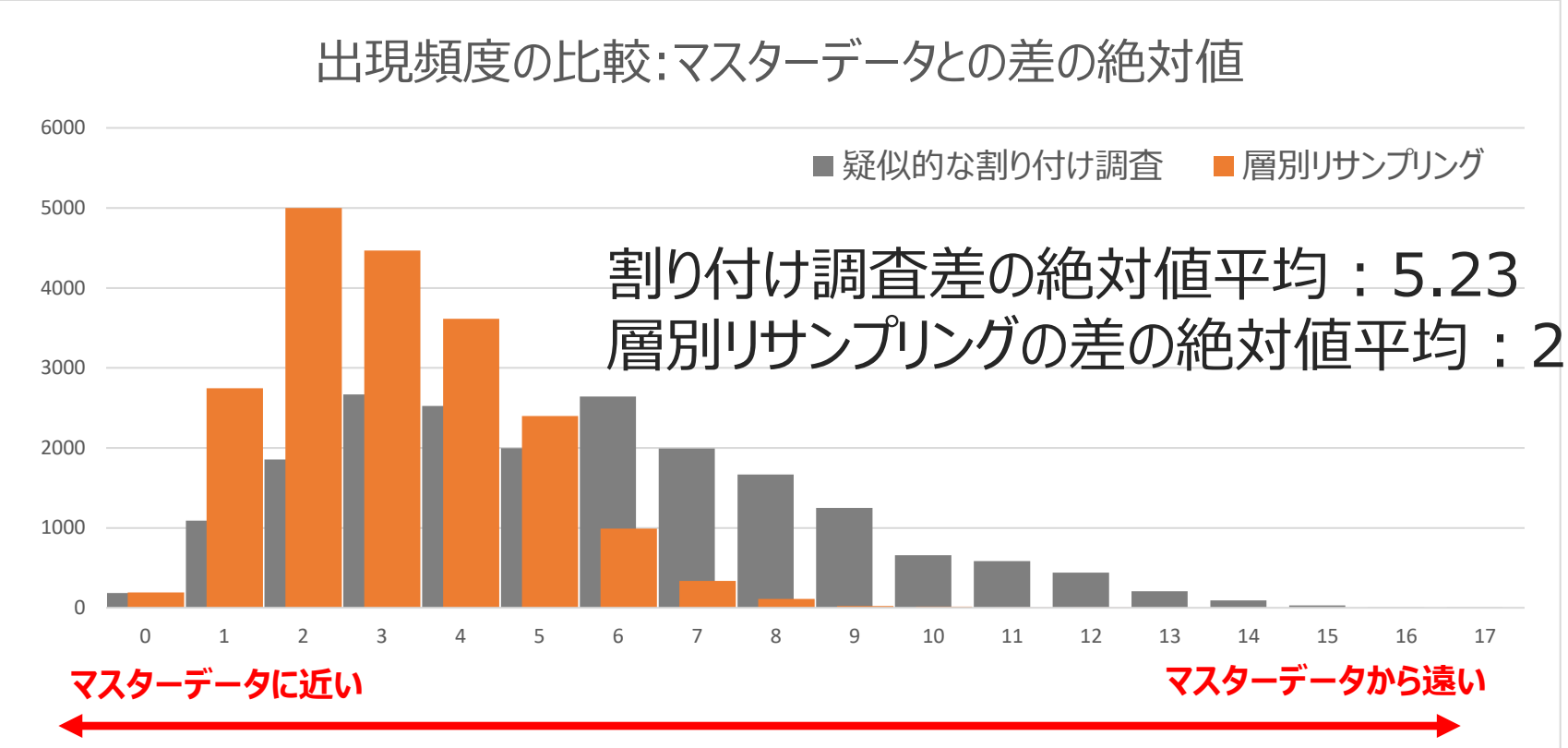
マスターデータとの差（マスターデータ - 層別リサンプリング）		Q1	
		YES	NO
Q２	YES	+0.5	+0.6
	NO	-1.1	+0.0

差の絶対値計：2.2

- クロス集計４セルの差の絶対値計をすべての組み合わせで実施

クロス集計の比較

- 疑似的な割り付け調査と層別リサンプリングのクロス集計の
  マスターデータとの差の絶対値の出現頻度
- すべての組み合わせ（19,900クロス）で実施



モンテカルロ法による検証

- 層別リサンプリングはランダムに抽出するので、試行する毎にデータ行列の中身は変わる
- 層別リサンプリングの生成を100回試行したら、これまで見てきた評価は変わるのか

マスターデータ（母集団と想定）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
260	591	839	877	265	619	777	834

サンプリング（非復元抽出）

疑似的な割り付け調査（サンプルサイズ）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
100	100	100	100	100	100	100	100

リサンプリング（復元抽出）

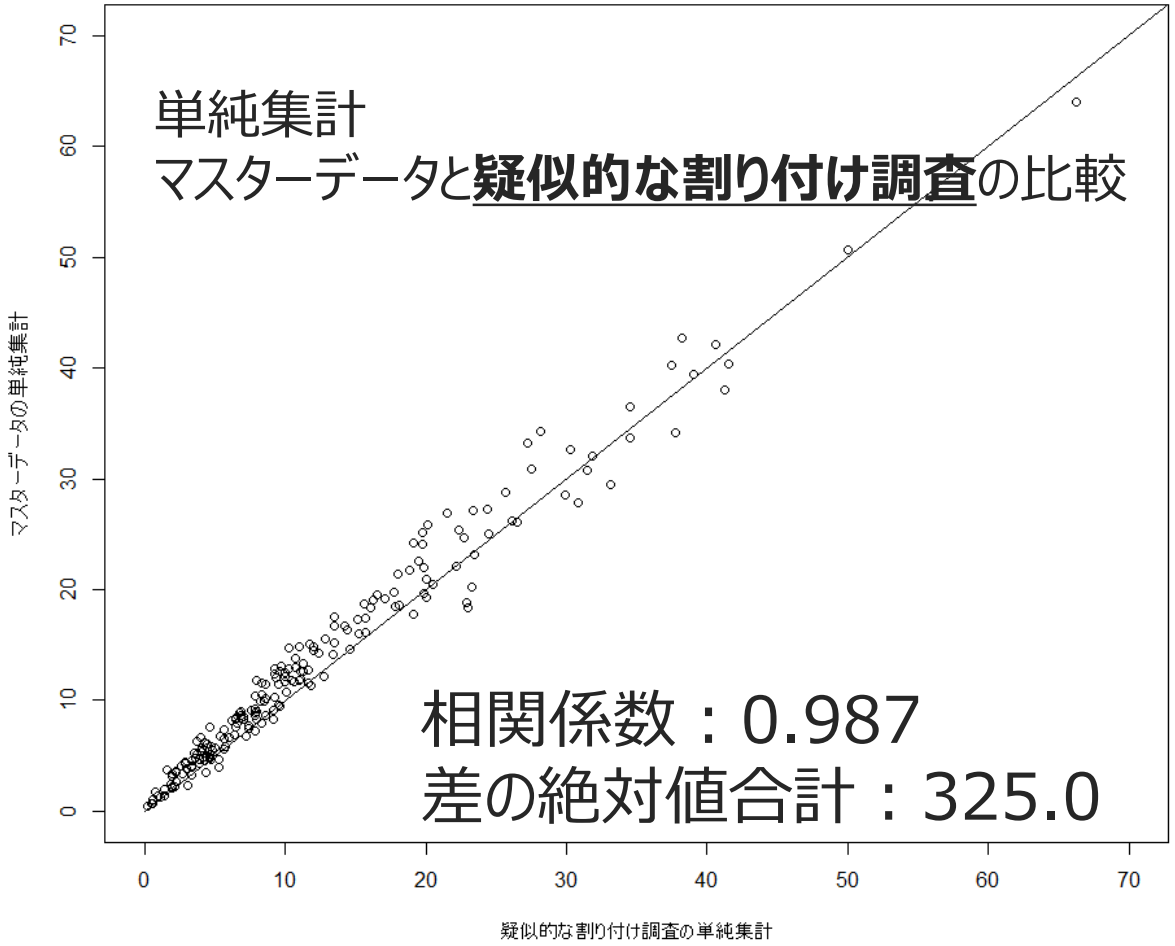
100回試行

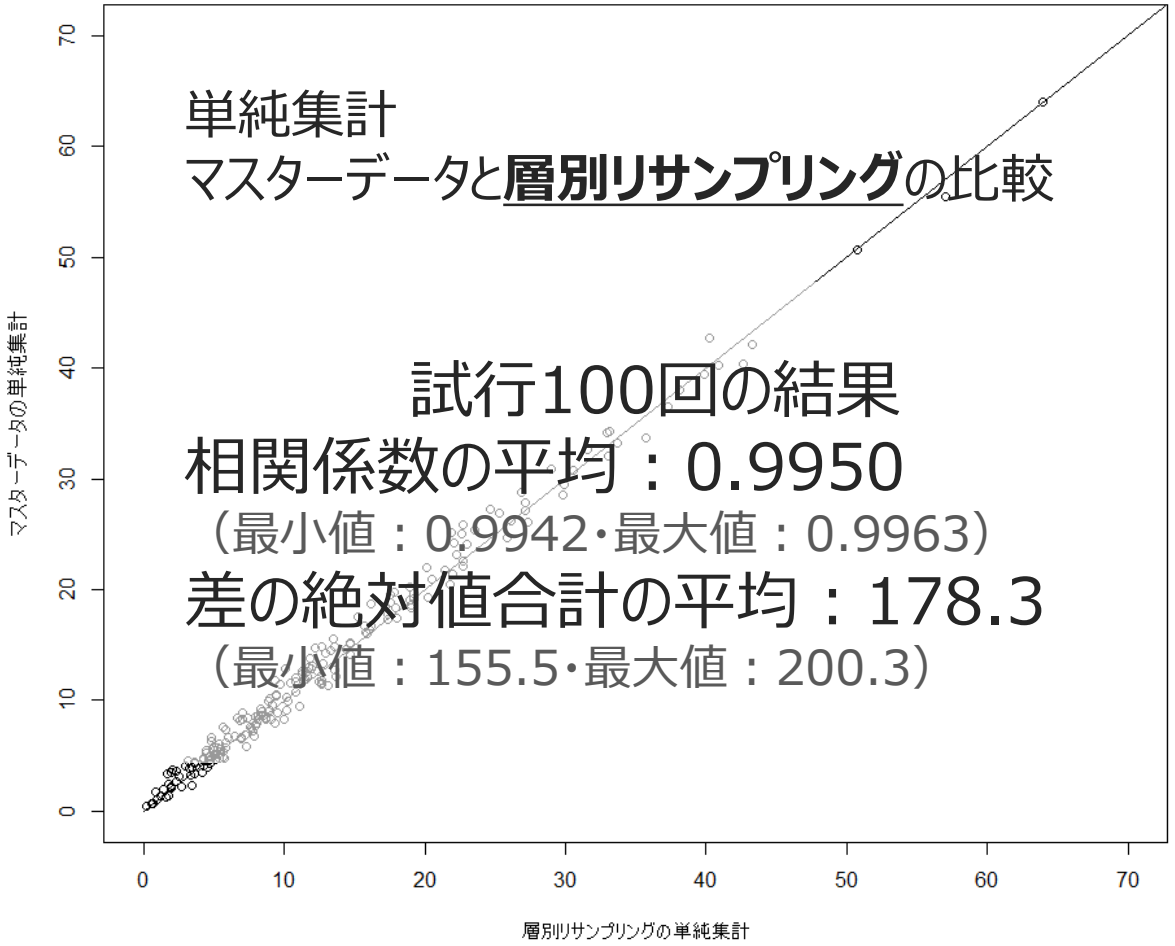
層別リサンプリング（サンプルサイズ）

男性				女性			
12-19才	20-34才	35-49才	50-69才	12-19才	20-34才	35-49才	50-69才
260	591	839	877	265	619	777	834

単純集計の比較

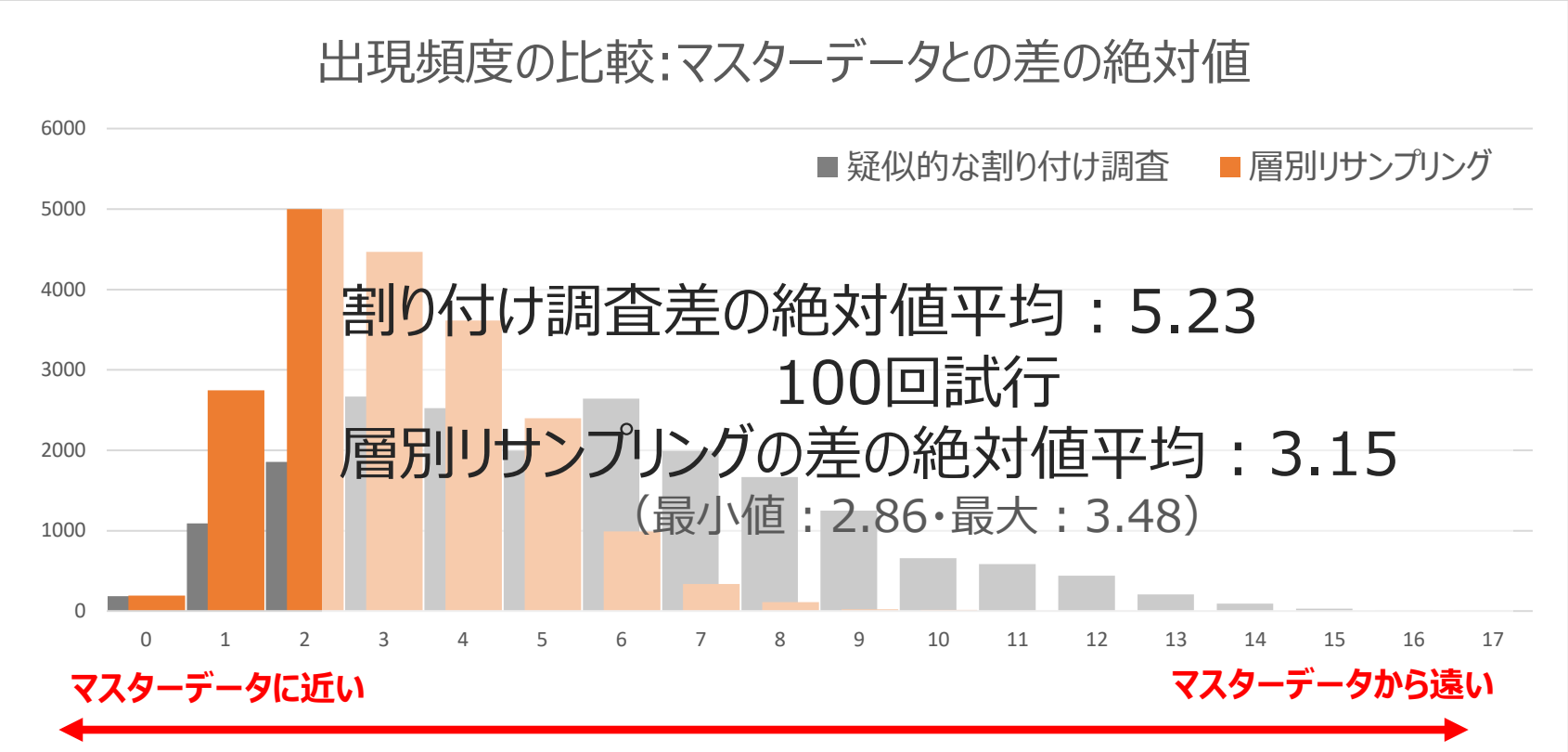
- 疑似的な割り付け調査と層別リサンプリングの単純集計結果を散布図で比較（割り付け調査を1回、層別リサンプリングを100回試行）
- 202項目をプロット





クロス集計の比較

- 疑似的な割り付け調査と層別リサンプリングのクロス集計のマスターデータとの差の絶対値の出現頻度（割り付け調査を1回、層別リサンプリングを100回試行）
- すべての組み合わせ（19,900クロス）で実施



ナビゲーターからの質問

① 同数でなく、任意の数に割り付けた調査データに対しても同じ補正が適用できるか

② クロス集計は2行2列でなく、任意のｍ行ｎ列のカテゴリーについても拡張できるか

③ 今回の補正データに推定と検定を適用できるか

相澤提案の意義は何か

1. 既存のブートストラップ法がｎ個のデータからｎ個のデータを復元抽出するのに対して、層別に異なるｍ個を復元抽出する方法である
2. 多変量解析でも抽出率の補正ができるようになった
3. 確率分布の仮定をおかないノンパラメトリックな方法なので適用範囲が広い

割り付け調査のデータを集計分析する対処法をぜひ教えてください

ご視聴ありがとうございました

課題への打開策は他にもいろいろあると思います。
ぜひ教えてください。意見交換しましょう。

「もっとうまいやり方がある」
「うちではこうしている」･･･　など、

ご意見・ご提言をJMRA宛にお送りください。

ご質問・ご意見は　⇒ office@jmra-net.or.jp

本キャストの過去動画は ⇒ https://www.jmra-net.or.jp/activities/seminar/2020/20210107.html

過去のご質問・ご意見は ⇒ https://www.jmra-net.or.jp/activities/seminar/2020/20210107v.html