惑わされない、怪しいデータの見分け方

『予想通りに不合理―行動経済学が明かす「あなたがそれを選ぶわけ」』（早川書房）などの著書で有名な行動経済学者、ダン・アリエリー氏が過去に共著で執筆した研究論文の一部に、データ捏造の疑いがかけられている。2012 年にアリエリー氏とその共著者は、不正を減らす方法についての研究を発表した[1]。そのなかで、「正直に回答することを約束する」署名の位置によって、回答者が正直に回答する割合が変わることを示した。

アリエリー氏らは、自動車保険会社が持つ顧客の自動車走行距離のデータを用いてフィールド実験を行った。顧客に所有する自動車の走行距離を申告させ、その回答用紙の最初に「正直に回答することを約束する」と署名するグループと、回答用紙の最後に同様の署名をするグループで顧客をランダムに割り振った。このフィールド実験では 2 万人に対し回答用紙を送付し、13,488 人から回答を得た。自動車の走行距離は前年も顧客に申告させており（前年は署名のグループは分けられておらず、全員同じフォーマットである）、現在の走行距離から前年の走行距離を差し引いて過去 1 年間における走行距離を算出した。

その結果、回答用紙の最初に署名したグループの過去 1 年の走行距離は、回答用紙の最後に署名したグループの走行距離よりも 2,400 マイル（10.25%）ほど長かった。走行距離が長いほど事故などのリスクが高まるため、顧客は走行距離を短くして報告するインセンティブがあると推測される。つまり、このフィールド実験は回答用紙の最初に署名することで、より正確な数値を回答者から引き出せることを示している。

しかし、後年に同様の実験をした際（Kristal et al. 2020）には、2012 年のような結果が得られなかった[2]。また、2012 年の自動車走行距離の実験のデータについて再度検証した結果、前年の走行距離について、最初に署名したグループより最後に署名したグループの走行距離がそもそも 15,000 マイルほど長くなっていた。グループをランダムに振り分けていたのであれば、このような大きな差は出ないはずである。そのため、Kristal et al.（2020）では、2012 年の論文（Shu et al. 2012）に掲載された自動車の走行距離に関するフィールド実験について、グループ間でのランダム化に失敗した可能性を指摘している。

その後、匿名の研究者たちによって自動車走行距離のデータに関する分析がさらに進み、こ

[1] Shu, L. L., Mazar, N., Gino, F., Ariely, D., and Bazerman, M. H.(2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. Proceedings of the National Academy of Sciences, 109, 15197–15200

[2] Kristal, A.S., Whillans, A.V., Bazerman, M.H., Gino, F., Shu, L.L., Mazar, N., Ariely, D., 2020. Signing at the beginning versus at the end does not decrease dishonesty. Proceedings of the National Academy of Sciences, 117, 7103-7107

のデータそのものが捏造であるとの疑義が呈された[3]。この匿名の研究者による分析では、データが捏造であることを裏付ける理由として、大きく分けて 3 つほど理由が掲載されている。その理由が非常にわかりやすく面白いため、このコラムで簡単に紹介したい。

1.　走行距離のデータ分布が異様である

自動車走行距離の分布は、正規分布のような山なりの分布を描くと考えられる。しかし、Shu et al(2012)のデータで分布を確認してみると、0 マイルから 50,000 マイルまでほぼ均一に分布していた。さらに 50,000 マイル以降の走行距離のデータは存在しなかった。

2.　走行距離下 3 ケタの数字がおかしい

自動車の走行距離などの数値を報告する際には、下 3 ケタの値などある程度数値を丸めて報告すると考えられる。実際、前年の走行距離のデータにおいては、下 3 ケタが「000」や「500」といった数値が多い傾向がみられた。一方、1 年後のデータで下 3 ケタの分布をみると、「000」から「999」が一様に分布していた。

3.　同じデータの中に異なるフォントがある（双子データ）

前回の自動車走行距離のデータのフォントが「Cambria」「Calibri」の 2 種類に分かれていた（元データは Excel ブック形式で公開されている）。「Cambria」「Calibri」それぞれのデータはほぼ同数であり、また、両データの走行距離の累積分布関数を見比べるとほぼ一致していた。さらに、「Cambria」のデータの下 3 ケタは一様に分布していた一方、「Calibri」のデータでは「000」「500」の数値が多い傾向がみられた。つまり、もともと存在していた「Calibri」のデータを複製し、そのデータに乱数を追加して「Cambria」データを捏造したことが示唆される。

データ分析を行う前に、そのデータの分布や統計量を確認することは非常に重要である。しかし、今回の件が示唆しているように、データを扱う際には統計的な部分に目を向けるだけでなく、どのようにそのデータが作成されたのか、そのプロセスなども含め細かな部分にも目を配る必要があろう。

（しめ鯖）

[3] 以下のサイトでデータ捏造の疑義に関する詳細や、実際に Shu et al. (2012)で使われたデータセットが公開されている。また、今回のデータ捏造の疑義について、アリエリー氏および共著者からのコメントも掲載されている。（http://datacolada.org/98）