

# 統計アラカルト

熊本の統計情報 令和4年12月27日

県民の皆様に統計を身近に感じていただくためのページです。

随時、色々な統計に関する話題・データを紹介します。

## データは嘘をつかない？ ～統計とバイアス～

現代社会において、「資料」はあらゆる場面で作られ、利用されていますが、「資料」の客観性・妥当性などを担保するものとして、多くの場合に「統計データ」が用いられています。それは、そこに記載された「データ」というものが客観的なものであると信じられており、「データは嘘をつかない」と思われているからです。

しかし、使用者がデータの使い方を間違ってしまうと、データの客観性が失われたり、データが「嘘をついた」ように見えたりすることがあるのです。いったいどういうことでしょうか？有名な事例を2つほど見ていくことにしましょう。

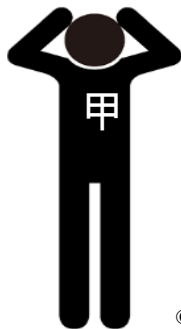
まず一つ目。「**データは正しいけれど、データの集め方に問題がある**」事例です。

[例] 甲学校と乙学校が学力を調査するため、生徒各100人に、選択科目A・Bのいずれかを選んでもらい、試験を実施しました。試験内容は両校で全く同じです。両校の科目ごとの平均点は以下の通りでした。

科目A、Bともに乙学校のほうが平均点が高いことが分かります。

	甲平均	乙平均
A	69 点	70 点
B	60 点	65 点

Lose



©dak

Win



©dak

ではこの時、2科目合計点の平均はどうなるでしょうか。調べたところ、なんと以下のような結果となりました！科目ごとでは平均点の低かった甲学校が、合計の平均点では乙学校を上回りました。

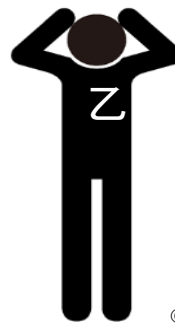
	甲平均	乙平均
A	69 点	70 点
B	60 点	65 点
合計	68 点	66 点

Win



©dak

Lose



©dak

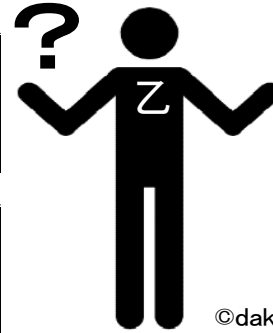
あれ？何か変な感じがしますよね。

一見すると、結果が逆転したようにも見えます。どうしてこのような現象が起きたのでしょうか。実は、この試験の各学校における科目A、Bの受験者数と合計点数は以下の通りでした。



〈選択者数〉

	甲平均	乙平均
A	90 人	10 人
B	10 人	90 人
合計	100 人	100 人



〈合計点数〉

	甲平均	乙平均
A	6210 点	700 点
B	600 点	5850 点
合計	6810 点	6550 点

©dak

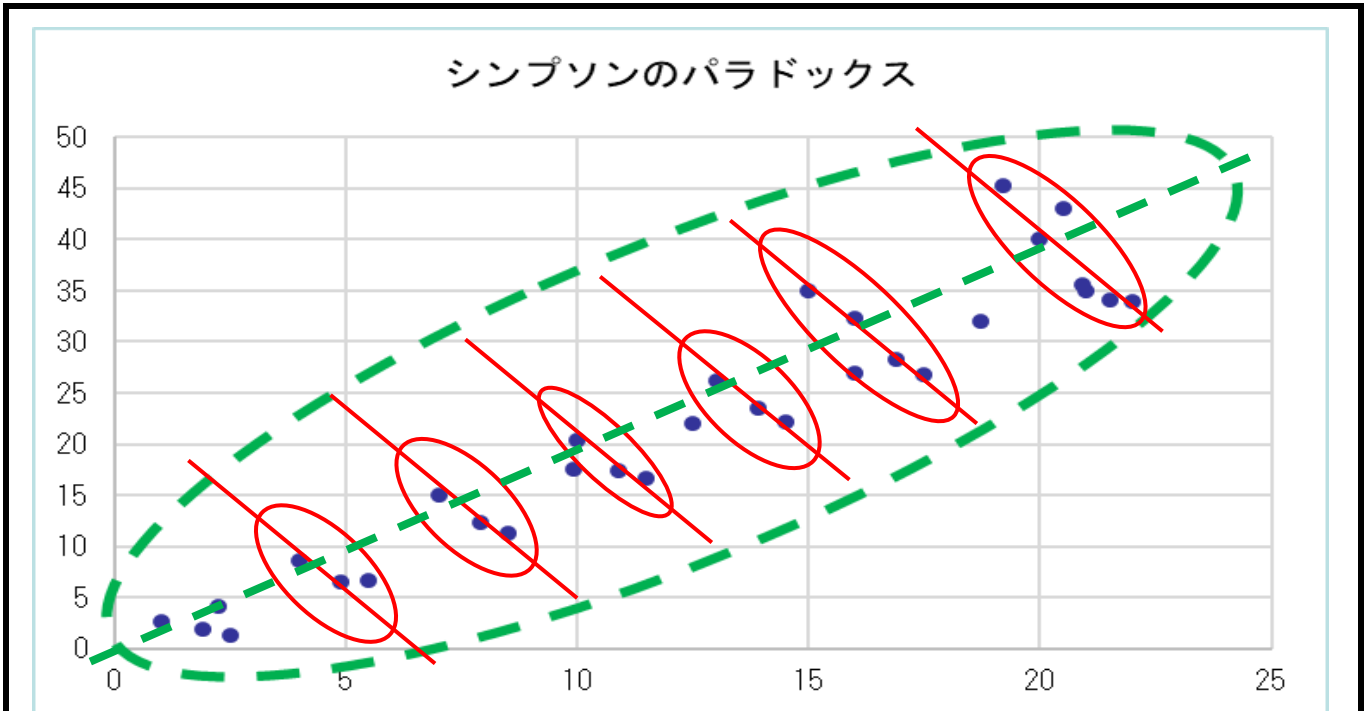
©dak

今回のテストでは、甲・乙両校の科目A・Bの受験者数に大きな偏りがあるため、調査対象の全体(合計)と部分(科目ごと)で逆の結果が生じています。

計算上の誤りはなく、その意味では「データは嘘をついていない」のですが、甲、乙両校の成績を比べるための客観的な資料としては望ましいとは言えません。

このように、調査の母集団に大きな偏り(バイアス)がある場合などに『全体と部分で一見すると矛盾したように見える結果が生じること』を、イギリスの統計学者エドワード・E・シンプソンにちなんで、「シンプソンのパラドックス」と呼びます。

「シンプソンのパラドックス」では、「個々で成り立つことが、全体で成り立つとは限らない」とされます。



図はシンプソンのパラドックスが生じているデータの分布をプロットしたもの。

○で囲われた部分のデータには負の相関(x軸が増加するとy軸は減少、右下がりの直線で図示)になっているが、○で囲われたデータ全体としては、正の相関(x軸が増加するとy軸も増加、右上がりの直線...で図示)になっている。

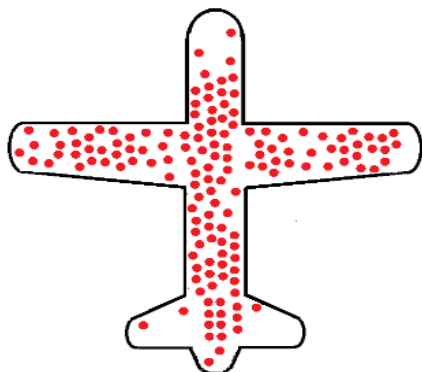
次に二つ目。「データは正しいけれど、人間の解釈が間違っている」事例です。

第2次世界大戦時、アメリカ合衆国が欧州に参戦した当初の米軍航空部隊では、軍用機の装甲が薄く、一説には隊員の生還率が50%であったと言われており、士気にも影響が出ていました。

そこで軍では、生還率を高めるため、機体の装甲強化を図ることにしました。しかし、航空機には重量制限があり、一部分しか装甲を強化することができません。

では一体、機体のどの部分の装甲を優先的に強化すればよいのでしょうか？

その優先付けをするため、軍は生還した機体の詳細なデータを集め、分析を行いました。

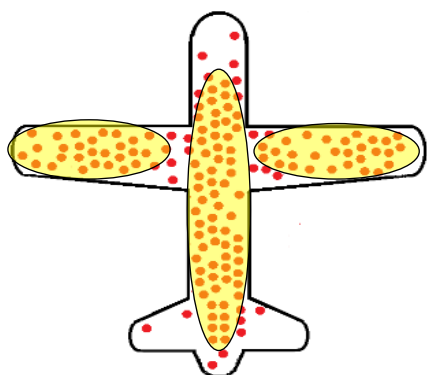


機体の損傷箇所プロット平面図例(イメージ)

分析の結果、損傷には明確なパターンがあることが分かりました。

生還したどの機体も、「翼と胴体」が蜂の巣のように穴だらけになっていた一方で、コックピットなどほとんど損害が無かったところもあったと言われています。

そこで、軍ではこれらのデータから判断して、損傷のひどい「翼と胴体」の装甲強化を最優先にすべき、と考えました。



● 損傷の多い箇所(イメージ)

皆さんはどう思われますか？

データから見ると、一見、軍の判断は正しそうですけど…。

しかし、軍のこの判断に反対した一人の天才数学者がいました。

それは、コロンビア大学統計研究グループに所属して戦争計画に参加していたエイブラハム・ウォールド博士です。

軍の調査は「生還できた機体」のデータのみ収集されたものであって、「生還できなかった機体」についての考察が抜け落ちておりと指摘しました。ウォールド博士は軍のデータを基に、生還できなかった機体を含めた全機体の損傷箇所の分布を推定します。

そして、「翼や胴体」は損傷しても生還できる箇所であり、データ上は損傷の少ない「エンジンとコックピット」こそがそこを撃たれたら生還できないウィーク・ポイントであると看破しました。

『…仮説データのこの分析は、被弾場所ではエンジンエリア、被弾の種類では20mm機関砲に対して航空機が最も脆弱であるという結論となる。

航空機がエンジンエリアに20mm砲弾を被弾した場合、被撃破率は0.534で最大となる。次に脆弱なのは、コックピットへの7.9mm機関銃による攻撃である。

…導出された脆弱性一覧表から得られる他の結論は、防護装甲の位置を特定するためのガイドとして使用でき、将来の作戦の推定損失の予測に使用できる。…』(傍線引用者)

Wald, Abraham. (1943). A Method of Estimating Plane Vulnerability Based on Damage of Survivors. Statistical Research Group, Columbia University. CRC 432 — reprint from July 1980. Center for Naval Analyses.より一部抜粋し、原文(英文)を和訳しました。

最終的に米軍は、ウォールド博士の提案を取り入れ、生還率を向上させることに成功しました。

この事例は「**生存(者)バイアス**(英語: survivorship bias、survival bias)」の好例として知られており、「**何らかの選択過程を通過した人・物・事のみを基準として判断を行い、その結果には該当しない人・物・事が見えなくなる**こと」(Wikipedia「生存者バイアス」より引用)を言います。

この事例以外の生存者バイアスとしては、企業の業績を調査する際に倒産企業からのデータが含まれないために、分析結果が過度に楽観的なものになる、などがあり、誤判断に繋がる危険性が指摘されています。

**バイアス(bias)**はいろんな局面で使用される言葉ですが、統計学用語では「**偏り**」と訳され、「母集団の要素が標本として平等に選ばれていない、または推定すべき量を何らかの理由で高く、または低く推定しすぎていること」(Wikipedia「バイアス」より引用)とされています。

**どれだけ正確なデータを用いて統計分析を行っても、何らかのバイアスによって調査対象を正しく掴んでいなければ、誤った結果しかもたらされません。**

バイアスに気を付けることは、ビッグデータなど詳細かつ大量のデータ分析ができるようになった、現代社会でこそ重要なのかもしれません。

問合せ先: 熊本県企画振興部 交通政策・統計局 統計調査課 総務資料班 〒862-8570 熊本市中央区水前寺6-18-1

電話: 096-333-2174 / Fax: 096-384-7544 / メール: toukeichousa@pref.kumamoto.lg.jp