データサイエンスへの誘い

第9回: 統計のウソ

瓜生真也（デザイン型AI教育研究センター・助教）

今日の目標

データ分析を行う上で留意すべき

統計のウソや誤用を知る

week09/09_reproducible_research.ipynb

誤差とバイアス

データ分析の際の前提条件として

誤差　さまざまな要因によって引き起こされる真の値との差

真の値 = 観測値 + 誤差

観測機、観測者、観測環境などに依存して発生する誤差→測定誤差（偶然誤差）

バイアス（系統誤差）　真実を歪ませる情報の偏りや考え方の総称

- 選択バイアス
- 測定バイアス
- 情報バイアス

不都合はすべて誤差のせい？

地上の高台から観測された惑星や彗星の位置と、予測された位置とは厳密には一致しなかった。ラプラスとその同僚たちは、このことを観測上の誤りのせいにした。（中略）ラプラスはこれらの誤りをすべて、数学的描写を付加した特別なもの（誤差関数）として片付けた。

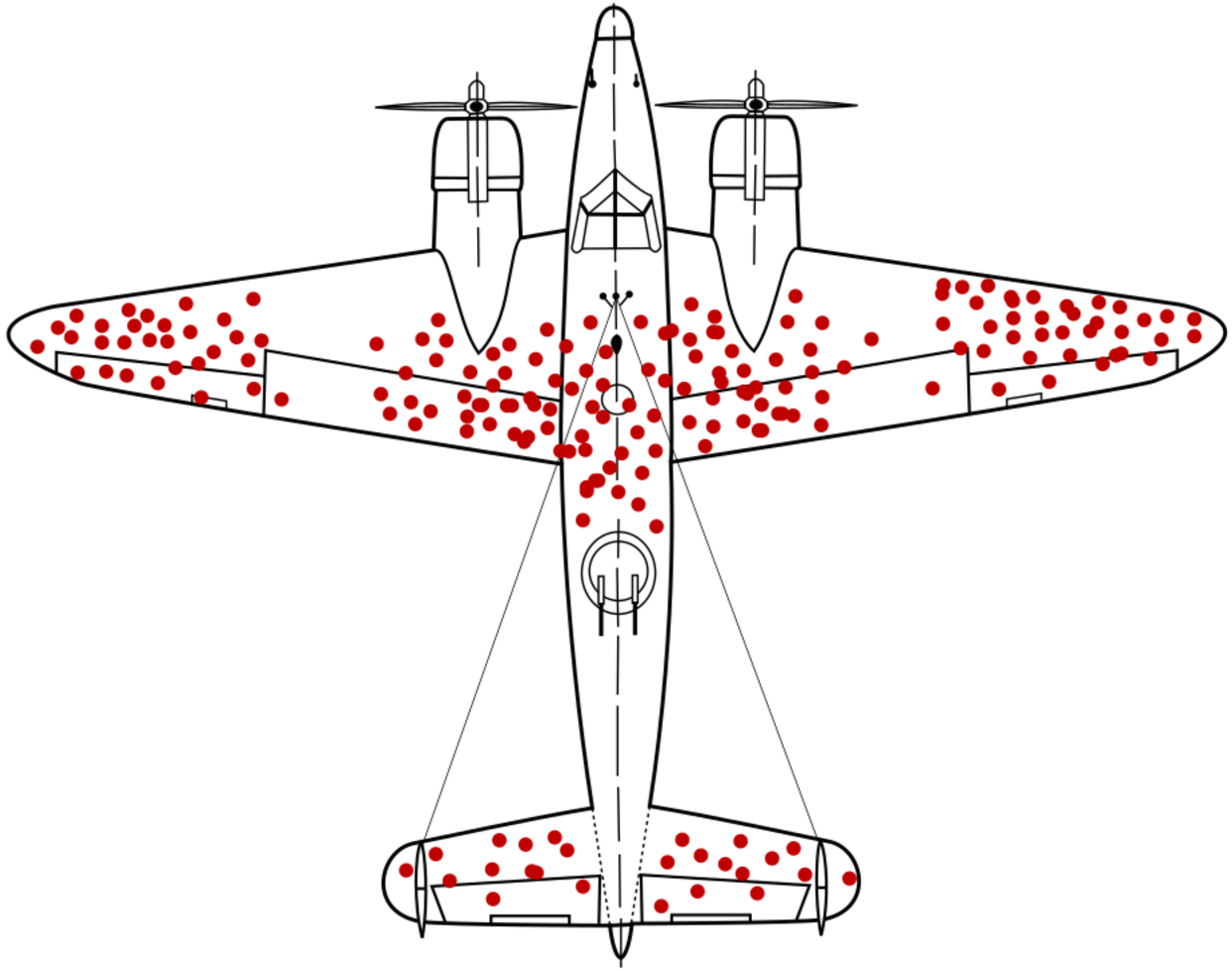
デイヴィッド・サルツブルグ 著、竹内惠行 訳、熊谷悦生 訳 「統計学を拓いた異才たち」
日本経済新聞出版社 (2010)より引用

エイブラハム・ウォールドの生存者バイアス

第二次世界大戦中、任務から戻った
機体について、損傷箇所を分析

赤い丸が損傷箇所

どこを補強するのが適切だろうか



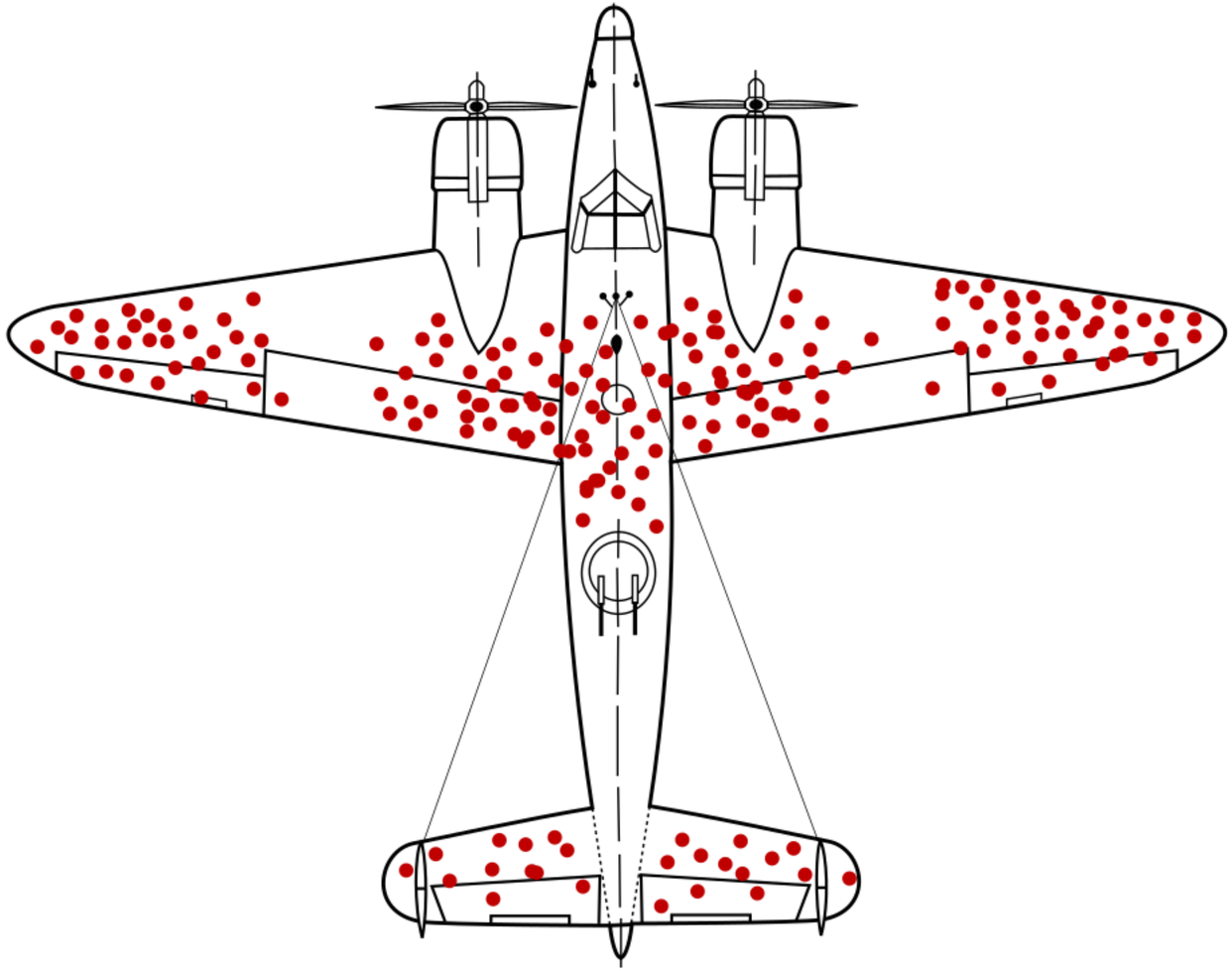
エイブラハム・ウォールドの生存者バイアス

ウォールドは撃墜された爆撃機が分析に含まれていないことを指摘

→生存したものだけを分析

帰還した機体が損傷を受けていない箇所を補強するように指示

赤い丸で示す箇所は損傷を受けても安全に帰還できる場所として考えたもの



第一種の過誤・第二種の過誤

仮説検定を行う際、誤った判断を引き出す原因となる2種類の誤り

仮説検定の結果	真実	
帰無仮説を棄却	第一種の過誤	正しい解釈
帰無仮説を採択	正しい解釈	第二種の過誤

対策例…　有意水準を小さく設定する
標本サイズを大きくする
検出力を評価する
トレードオフの関係にあるので注意

意図せずに結果を間違えて導き出す可能性もある

week09/09_reproducible_research.ipynb

p値の誤解と悪用

仮説検定におけるp値の誤解

p値に対する間違った解釈

✕ 帰無仮説が真である確率、正しい可能性を示す

✕ 得られたデータが偶然の結果である可能性を示す

✕ 1 - pにより、得られたデータの確からしさを示す

✕ 有意水準未満ならば結果は価値がある

○ 帰無仮説が真であるという前提の下で、
想定する統計モデルが正しく、
データにバイアスが含まれずランダムに得られている場合において
観測されたデータ以上に極端な結果が得られる確率

p値の悪用: p-hacking

都合の良いようにp値を操作する

p値は

1. （差がないとする）帰無仮説と得られたデータとの違いの大きさ
2. **サンプル数**

に依存する

サンプル数を増やすことでp値の操作が可能

対策　信頼区間や効果の大きさなど他の統計的な指標も含めた総合的な評価
　　ベイズ的アプローチの利用

week09/09_reproducible_research.ipynb

再現性

再現性（Reproducibility）

同じデータや手法を用いることで、一貫性のある結果が得られる性質

再現性が高い　誰がいつ、どこで再現をしても同じ結果となる

→分析結果の信頼性・透明性を高め、意思決定に効果的な結果となる

再現性が低い　他者が異なる環境・時間で行うと、同じ結果を得られない

→分析結果そのものの信頼性をなくす、偶然的な結果とみなされる可能性

プログラミングを用いたデータ分析では

コード　文章　結果　図表

を一箇所で管理することが重要

バージョン管理システムの利用を推奨