

# 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)

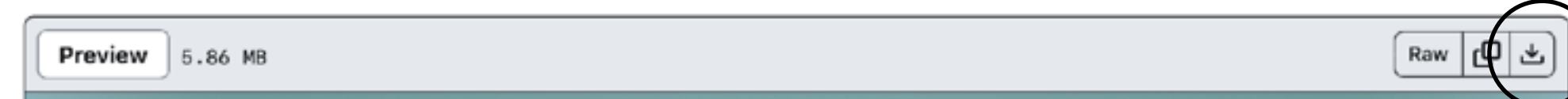
# 講義内容（予定）

講義に関する資料（スライド、補足資料等）を  GitHubに置いておきます

 <https://github.com/uribo/INNV1250>



ダウンロード可能



1. ガイダンス、データサイエンスとは何か
2. 現代社会におけるデータサイエンスの活用事例
3. データ処理の手法
4. データの要約
5. データの可視化
6. データと確率
7. データからの推論
8. 複数のデータを比較する

9. 統計のウソ
10. 統計モデリング
11. 統計的学習 ゲスト講師による特別講演を企画中
- ~~12. さまざまなデータサイエンスの手法~~
13. 機械学習とAI
- ~~14. コンピューターを用いた分析~~
15. ビッグデータの扱い
16. 期末試験（8月1日）

# 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

# 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

$$\text{真の値} = \text{観測値} + \text{誤差}$$

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。

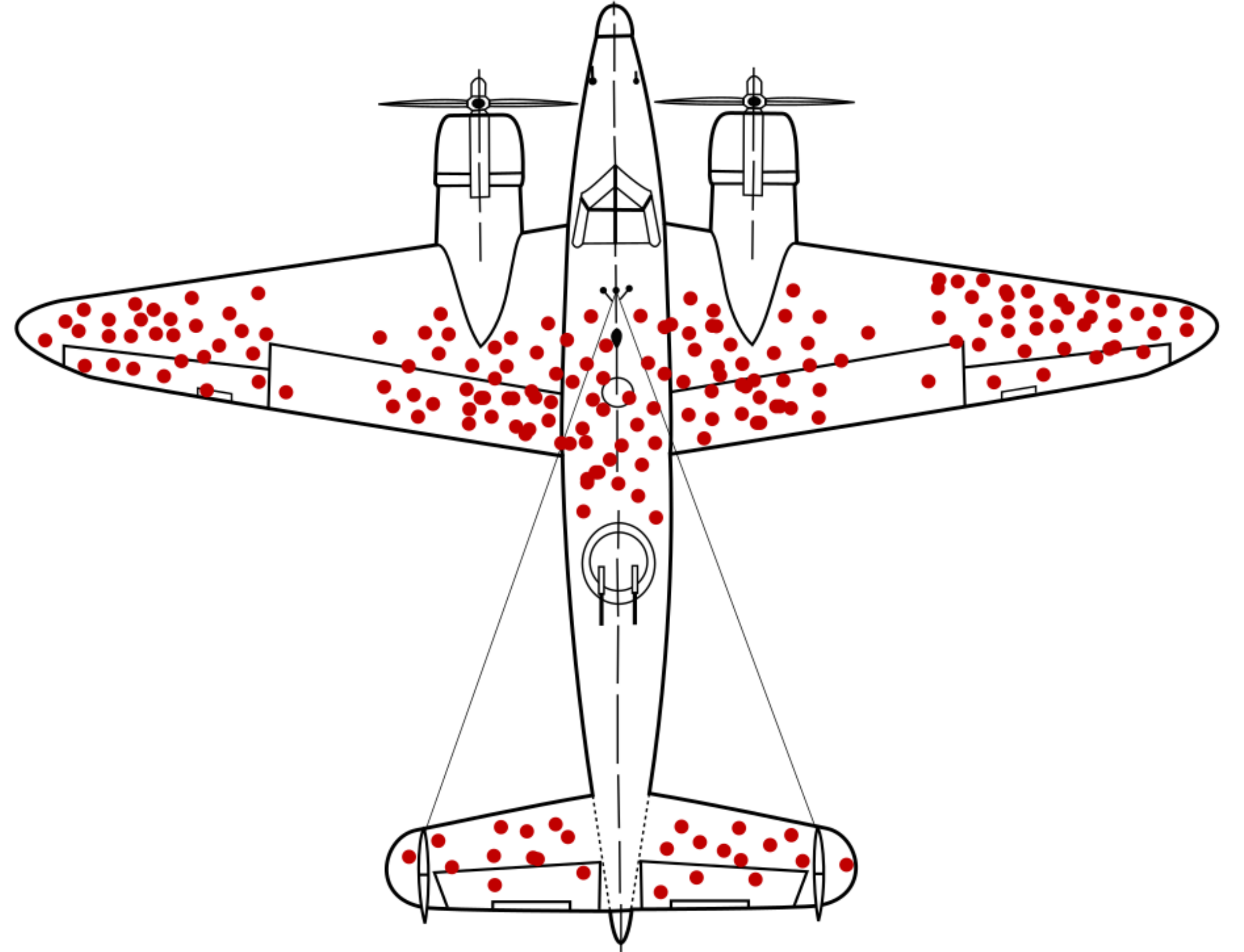
デイヴィッド・サルツブルグ 著、竹内恵行 訳、熊谷悦生 訳 「統計学を拓いた異才たち」  
日本経済新聞出版社（2010）より引用

# 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



Martin Grandjean (vector), McGeddon (picture), Cameron Moll (concept)

CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>

Wikimedia Commonsより

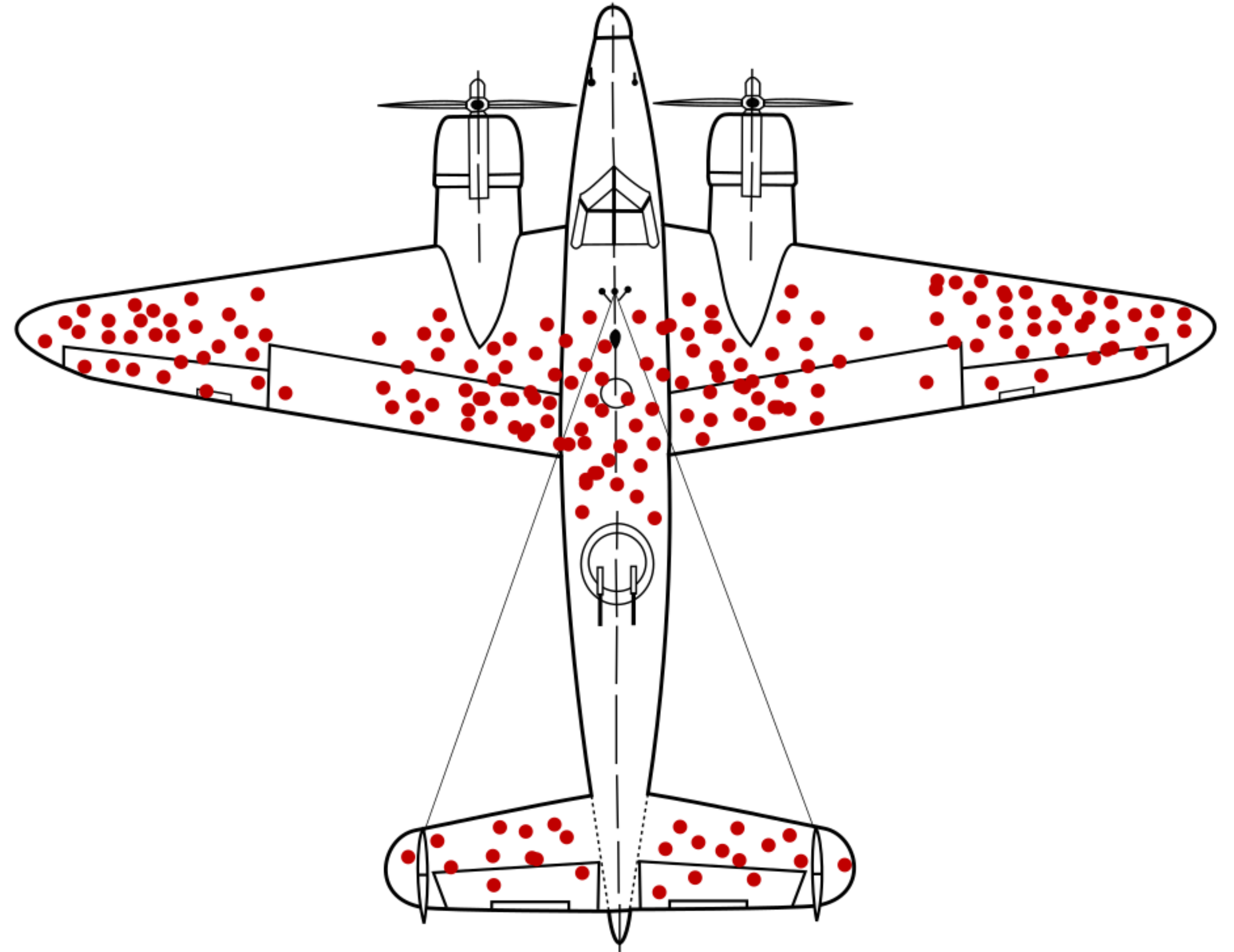


# エイブラハム・ウォールドの生存者バイアス

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  
→生存したものだけ进行分析

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



Martin Grandjean (vector), McGeddon (picture), Cameron Moll (concept)

CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>

Wikimedia Commonsより

# 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	
	帰無仮説が正しい	帰無仮説が間違い
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



# 現実問題としての第一種の過誤・第二種の過誤



## 火災報知器の振る舞い

火事が起こっていない現場

  報知器が作動（誤作動）→「ない」ものをあると判断してしまう

第一種の過誤

火事が起こっている現場

  報知器が作動せず→あるものを「ない」と判断してしまう

第二種の過誤

真実を見落として、意思決定が行えない（この場合は消火活動など）ため、  
**第二種の過誤**を起こすことが問題となる

状況に応じて問題となる過ちの種類は異なる

# 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

# 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



# 再現性研究のためのJupyterの利用

# Jupyterで扱うipynb形式のファイル（テキストファイルの一種）

### 相関分析: 統計的手法を用いた比較

データの比較・関係を把握するためには、統計的手法を用いることもできます。ここでは、相関分析を紹介します。

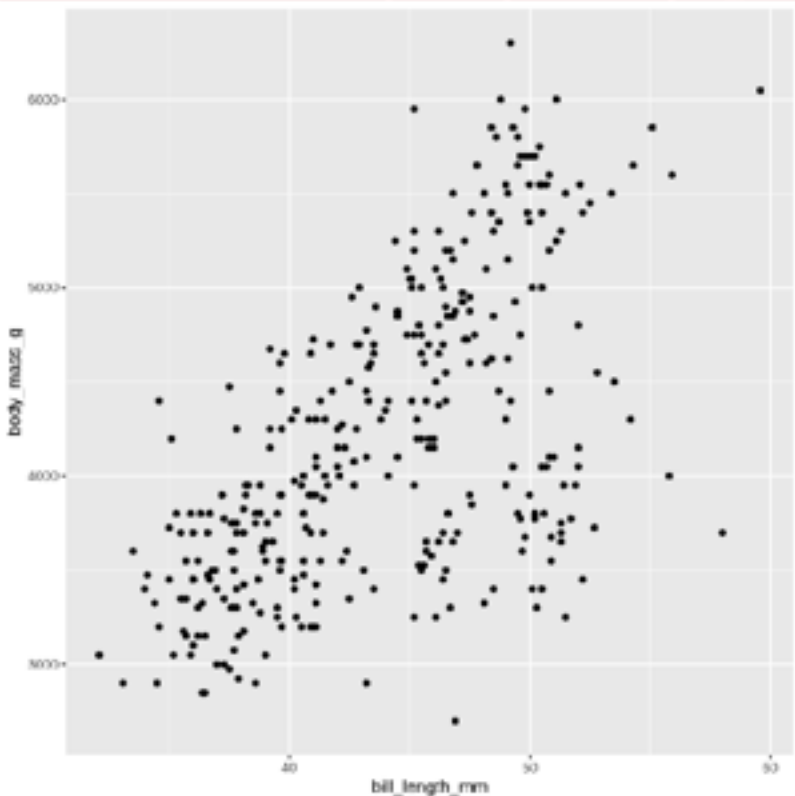
相関分析は、2つの変数間の関係を数値化する手法です。例えば、南極に生育するペンギンのくちばしの長さと体重の関係を調べた際、次の散布図から、くちばしが長い個体では体重もはわかりません。そこで変数間の関係性の程度を示す統計量を求めることで、その関連性を評価できるようになります。

```
[2]: # ペンギンデータの読み込み
penguins <-
  readr::read_csv("https://raw.githubusercontent.com/allisonhorst/palmerpenguins/main/inst/extdata/penguins.csv",
    col_types = "ccdddddcd")

ggplot(data = penguins) +
  aes(bill_length_mm, body_mass_g) +
  geom_point()
```

コード

```
Warning message:
"Removed 2 rows containing missing values ('geom_point()')."
```



## 結果 図表

- 共分散 (covariance)

共分散は、2つの変数間の関係を数値化する手法の一つです。共分散は、2つの変数の偏差積の平均値を表現します。

以下の式で2つの変数の間の共分散を求めることができます。

$$\begin{aligned}\text{cov}(X, Y) &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}\end{aligned}$$

文章

コード

ファイルの閲覧・編集に  
特別なソフトウェアは不要

```
"attachments": [],
"cell_type": "markdown",
"id": "a7a54acd",
"metadata": {},
"source": [
"# 相関分析: 統計的手法を用いた比較\n",
"\n",
"# データの比較: 関係を把握するためには、統計的手法を用いることもできます。ここでは、相関分析を紹介します。 \n",
"\n",
"# 相関分析は、2つの変数間の関係を数値化する手法です。例えば、南極に生育するペンギンのくちばしの長さや体重の関係を探った際、次の散布図から、くちばしが長い個体では体重も大きい傾向にあることがわかります。一方、どの程度の関連性があるのかといったことは散布図だけではわかりません。そこで変数間の関係性の程度を示す統計量を求めることで、その関連性を評価できるようになります。 \n"
],
},
{
"cell_type": "code",
"execution_count": null,
"id": "50196c4d",
"metadata": {
"vscode": {
"languageId": "r"
}
},
"outputs": [],
"source": [
"# ペンギンデータの読み込み\n",
"%perquins <- \n",
"% read::read_csv(\"https://raw.githubusercontent.com/allisonhorst/palmerpenguins/main/inst/extdata/penguins.csv\")\n",
"% col_types = \"c\" \n",
"\n",
"% ggplot(data = perquins) ~\n",
"% aes(bill_length_mm, body_mass_g) ~ \n",
"% geom_point()
],
},
{
"attachments": {},
"cell_type": "markdown",
"id": "9389720a",
"metadata": {},
"source": [
"# 共分散 (covariance)\n",
"\n",
"# 共分散は、2つの変数間の関係を数値化する手法の一つです。共分散は、2つの変数の関連性の平均値を表します。 \n",
"\n",
"# 以下の式で2つの変数の間の共分散を求めることができます。 \n",
"\n",
"
$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$


```

結果だけを共有するときは  
HTMLやPDFでの出力も候補となる

※ ノートブックとしての見栄えはJupyter環境で再現可能

# 参考資料・URL

目録 デイヴィッド・サルツブルグ(著), 竹内恵行・ 濱田悦生(訳)  
『「誤差」「大間違い」「ウソ」を見分ける統計学』（2021）  
東京大学出版会. ISBN: 978-4-320-11450-0

瓜生居室: あり、徳大図書館: あり、市立図書館: あり、県立図書館: あり

目録 阿部真人『統計学入門：データ分析に必須の知識・考え方』（2021）ソシム.  
ISBN: 978-4-8026-1319-4

瓜生居室: あり（電子版）、徳大図書館: あり、市立図書館: なし、県立図書館: なし

目録 竹内薫 『統計の9割はウソ：世界にはびこる「数字トリック」を見破る技術』（2004）  
徳間書店. ISBN: 978-4-19-863706-4

瓜生居室: なし、徳大図書館: あり、市立図書館: あり、県立図書館: なし

目録 Andrew Vickers (著), 竹内正弘 (監訳)  
『p値とは何か：統計を少しずつ理解する34章』  
(2013) 丸善出版. ISBN: 978-4-621-08551-6

瓜生居室: なし、徳大図書館: あり、市立図書館: なし、県立図書館: なし

目録 永田靖『サンプルサイズの決め方』（2003）朝倉書店.  
ISBN: 4-254-12665-4

瓜生居室: あり、徳大図書館: あり、市立図書館: なし、県立図書館: あり

