付 録 B

本書で利用したデータ

本書で利用したデータはこの付論で説明するが, 各データは次の場所から利用できる. **openintro.org/data**. (訳注 : https://www.jstat.or.jp/ 日本統計協会のウエブページからもダウンロードできる.) さらにこの付論には追加のデータセットがあるがこれはさらに勉学の成果に磨きをかけるためのものである. 各データについて次のような情報を含んでいる.

- データセットの変数リスト
- CSV ダウンロード形式
- R プログラム

B.1 Chapter 1

1.1 データ`!stent30`, データ`!stent365` → データ Stent は二つのデータセット 0-30 日データと 0-365 日データに分けられている.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for ntracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
NY タイムズ論説: www.nytimes.com/2011/09/08/health/research/08stent.html.

1.2 データ`!loan50`, データ`!loans_full_schema` → このデータは貸付組合 (Lending Club), (lendingclub.com), が提供, この組合を通じてローンを受けた人々の大きなデータがある. テキストで利用したのは第一四半期 (1 月 2 月, 3 月), 2018 年のサンプルの一部分である.

1.2 データ`!county`, データ`!county_complete` → このデータは幾つかの公的統計からとったものである. データ county に含まれる変数については 2018 年後半に利用可能であった近年のデータを利用した。より最新のデータの出所は以下のとおり, census.gov, (ただし 2021 年現在ではデータをとった Quick Facts ページは今は利用可能ではない.) 最近のデータの出所は以下である.
USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), American Community Survey (census.gov/programs-surveys/acs).

1.3 データ Nurses' Health Study → このデータ (看護師健康調査) について詳しくは以下に情報がある: www.channing.harvard.edu/nhs

1.4 単純ランダム化 (ブロック化なし) 研究の議論で念頭にあったのは次の研究である.
Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

B.2 Chapter 2

2.1 データ`!loan50`, データ`!county` → このデータは第 1 章のデータと同一.

2.2 データ`!loan50`, データ`!county` → このデータは第 1 章のデータと同一.

2.3 データ`!malaria` →, Lyke et al. 2017. PfSPZ ワクチンは strain-transcending-T 細胞により人間のマラリア感染を予防する. PNAS 114(10):2711-2716.
www.pnas.org/content/114/10/2711

B.3 Chapter 3

3.1 データ`!loan50`, データ`!county` → このデータは第 1 章と同一.

3.1 データ`!playing_cards` → このデータはトランプ札 52 枚は標準的な 1 セット.

3.2 データ`!family_college` →, このデータはシミュレーションで生成したが, 次に挙げる実際の人口についての要約統計に基いている.
nces.ed.gov/pubs2001/2001126.pdf.

3.2 データ`!smallpox` →, Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

3.2 データ Mammogram screening, probabilities → このデータで報告した確率は次の研究で得られたものに基づく.
www.breastcancer.org, www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

3.2 データ Jose campus visits, probabilities → このデータは人工的に作成された.

3.3 この節で用いたデータはない.

3.4 データ Course material purchases and probabilities → このデータは人工的に作成された.

3.4 データ Auctions for TV and toaster → このデータは人工的に作成された.

3.4 データ`!stocks_18` → このデータは 2015 年 11 月〜2018 年 10 月の間のカタピラ (Caterpillar), エクソン (Exxon Mobil Corp), グーグル (Google) の株価の月次リターン.

3.5 データ`!fcid` → このデータは米国人口センサスからの単純なランダムサンプル (標本) と見なせる. 原データは USDA(Food Commodity Intake Database) による.

B.4 Chapter 4

4.1 データ SAT・ACT スコア分布 →SAT スコアデータ 2018 年の分布は以下で公表されている
reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf
ACT スコアデータは以下で利用可能
act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.pdf
実際の ACT スコアの分布は正規分布に近くないことを認識する必要がある. この話題の原データは簡単に手に入るので説明と例のみを用いることにした.

4.1 データ Male heights → このデータ (男性身長) で利用した分布は USDA の Food Commodity Intake データ・ベースに基づく.

4.1 データ`!possum` → このデータ (ポッサム) の分布のパラメターはオーストラリアとニューギニアのポッサムの標本データに基いている. このデータの出典は Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala).* Australian Journal of Zoology 43: 449-458.

4.2 データ Exceeding insurance deductible → このデータ (過剰な保険控除) の数値は作ったものであるが, 低控除プランでは観察され得る数値である.

4.3 データ Exceeding insurance deductible → この数値は作ったものであるが, 低控除プランでは観察され得る数値である.

4.3 データ Smoking friends → このデータ (喫煙) は残念ながら現在では 30%の統計のソースについての情報が失われ正確に証明できないので, この数値を事実と見なすことはできない.

4.3 データ US smoking rate → このデータ (喫煙率) は US の 15% 喫煙率でウェブサイト (Centers for Disease Control and Prevention) の値, 2017 年の 14%に近い.
cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

4.4 データ Football kicker → このデータは人工的に作成された.

4.4 データ Heart attack admissions → このデータは人工的だが, この心臓発作の認定は病院では現実的なはずである.

4.5 データ`!ami_occurrences` → このデータシミュレーション・データだが, 典型的な AMI 率はニューヨーク市の AMI(急性心筋梗塞) データに似ている.

B.5 Chapter 5

5.1 データ`!pew_energy_2018` → このデータの観察値はこの章で引用したよりも多い. ここでは副サンプル (sub-sample) を用いたが, 少し変動性が増すよう幾つかの例をスムーズにしている. データ pew_energy_2018 のフル・データセット, 異なるエネルギー源, 太陽光, 風力, 海洋掘削, 水力, 原子力などを含んでいる. データの作成に利用した統計は次のページからの引用.

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

5.2 データ`!pew_energy_2018` → このデータについて詳しくは 5.1 節のデータの説を参照.

5.2 データ`!ebola_survey` →, このデータは 2014 年 10 月 23 日ニューヨーク市, ギニアでエボラを治療していたある医師が熱がある病院に出かけたところエボラと診断されたた直後のもので, NBC 4 New York/The Wall Street Journal/Marist 世論調査で 82% のニューヨーク市民は「エボラ患者に接触したすべての人に 21 日間の強制隔離」に賛成と報じた. この世論調査は 2014 年 10 月 26 日-28 日に 1,042 名のニューヨークの成人からの回答である. Poll ID NY141026 on maristpoll.marist.edu.

5.3 データ`!pew_energy_2018` → このデータについて詳しくは 5.1 節のデータ説明を参照されたい.

5.3 データ Rosling questions → このデータはハンス・ロスリング (Roslings) が書籍で説明しているデータから小さい標本を採った. ここで述べている標本は同様だが実際の数値と同じではない. (時々は) 本で議論している母集団についての二つの質問への正しい回答での大体の比率は以下にある.

- 世界の 1 歳児の 80% は何らかの病気に対するワクチンが投与されている : 13% が正しい答え (米国では 17%).
- 2100 年の世界での子供の数 : 9% が正しい.

幾つかの追加の質問と正解を得るための大体の割合 :

- 今日の世界の低所得国において初等教育を終える女子児童: 20%, 40%, or 60%? 答え : 60%. 約 7%の人が正解を述べる.

- 今日の世界の平均寿命: 50 歳, 60 歳, あるいは 70 歳？答え： 70 歳. 米国では約 43% の人が正解を述べている.
- 1996 年にタイガー, ジャイアント・パンダ, クロサイは絶滅危惧にリストされている. これら 3 種の内, 今日に非常に絶滅が危惧されている種はなんだろうか. 2 つ, 1 つ, どれも異なる？答え： どれでもない. 約 7% の人々が正解が分かる.
- 世界の何パーセントが電気を利用できるだろうか？正解：80%. 約 22% の人々が正解に到達する.

より詳しい情報については文献を見られたい.

5.3 データ`!pew_energy_2018` → このデータについて詳しくは 5.1 節のデータ説明を参照されたい.

5.3 データ`!nuclear_survey` → このデータは 2013 年 3 月での 1,028 名の米国成人のランダムサンプル, 56% の米国成人が核兵器削減に賛成している.
www.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx

5.3 データ Car manufacturing → このデータは人工的に作成された.

5.3 データ`!stent30`, データ`!stent365` → このデータ第 1 章と同一.

B.6 Chapter 6

6.1 データ Payday loans → このデータは以下による.
pewtrusts.org/-/media/assets/2017/04/payday-loan-customers-want-more-protections-methodology.pdf

6.1 データ Tire factory → このデータは人工的に作成されたもの.

6.2 データ`!cpr` → Böttiger et al. *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial.* The Lancet, 2001.

6.2 データ`!fish_oil_18` → Manson JE, et al. 2018. *Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer.* NEJMoa1811403.

6.2 データ`!mammogram` → Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

6.2 データ`!drone_blades` → このデータはクワッドローター・ドローン翼の品質管理データで例のために作成した架空のデータ. シミュレーションをデータ`drone_blades` に示した.

6.3 データ`!jury` → このデータは人種差別審査の陪審員データで例のために作成したもの. シミュレーションデータをデータ`jury` に加えた.

6.3 データ`!sp500_1950_2018` → このデータのソースは以下である. finance.yahoo.com.

6.4 データ`!ask` → このデータは Minson JA, Ruedy NE, Schweitzer ME. **ばかげた質問もあるが戦略的コミュニケーションの一環**. .
opim.wharton.upenn.edu/DPlab/papers/workingPapers/
Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

6.4 データ`!diabetes2` → Zeitler P, et al. 2012. *A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes.* N Engl J Med.

B.7 Chapter 7

7.1 データ Risso's dolphins → このデータ (ハナゴンドウ・イルカ) は Endo T and Haraguchi K. 2009. *High mercury levels in hair samples from residents of Taiji, a Japanese whaling town.* Marine Pollution Bulletin 60(5):743-747. Taiji は映画 *The Cove* からであり, 日本におけるイルカとクジラ食についてのソースである. 毎年何千ものイルカが太地町 (Taiji 地域) を通過, 19 匹のイルカが多くのイルカのランダムサンプルを表現している.

7.1 データ Croaker white fish → fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

7.1 データ`!run17` → www.cherryblossom.org

7.2 データ`!textbooks`, データ`!ucla_textbooks_f18` → このデータは 2010 年と 2018 年に OpenIntro の関係者が集めたものである. 2018 年のサンプルは 201 の UCLA 授業からサンプルをとったが, 68 の教科書はアマゾンで見つけられる. ウエブサイトの情報は以下である. sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, amazon.com.

7.3 データ`!stem_cells` → Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012.

7.3 データ`!ncbirths` → このデータの出生記録は 2004 年にノースカロライナで発表された. 残念ながらこのデータセットについての追加的な情報は得られない.

7.3 データ Exam versions → このデータは人工的に作成された.

7.4 データ Blood pressure statistics → このデータは血圧 140-180 mmHg 間の患者の標準偏差は想像であり実際の観察データであるが (乖離は大きくはないとは云え) 少し不正確である.

7.5 データ`!toy_anova` → このデータは図 7.19 より作成された.

7.5 データ`!mlb_players_18` → mlb.mlb.com/stats からとられたデータ. 少なくとも 100 回バッティングした選手のみを分析で考慮されている.

7.5 データ`!classdata` → このデータは人工的に作成された.

B.8 Chapter 8

8.1 データ`!simulated_scatter` → 最初の 3 プロットでは架空のデータを使用した。完全直線フィットはグループ 4 データを用いたが、変数 `group` は図 8.1 のデータである. 3 つの不完全線形プロットはグループ 1-3 データを用いている (図 8.2). 三角関数カーブはグループ 5 データ (図 8.3). 残差を含めた 3 つの散布図はグループ 6-8 データによる (図 8.8). 相関プロットはグループ 9-19 データによる (図 8.9 と図 8.10).

8.1 データ`!possum` → このデータは第 4 章と同一.

8.2 データ`!elmhurst` → このデータはエルムスハースト (Elmshurst) 大学の 2011 年入学の 1 年生のデータからサンプルされたもので, 次の論文による. *What Students Really Pay to Go to College* オンラインで *The Chronicle of Higher Education* で公刊されている. chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

8.2 データ`!simulated_scatter` → 上手くいかないプロットにはグループ 20-23 データによる (図 8.12).

8.2 データ`!mariokart` → このデータは任天堂 Wii のマリオカートゲームの ネット (Ebay, ebay.com) オークションで 2009 年 10 月初旬に得られたもの.

8.3 データ`!simulated_scatter` → このデータの外れ値のタイプのプロットはグループ 24-29 による (図 8.18).

8.4 データ`!midterms_house` → このデータは Wikipedia からとられた

B.9 Chapter 9

9.1 データ`!loans_full_schema` → このデータは第 1 章と同一.

9.2 データ`!loans_full_schema` → このデータは第 1 章と同一.

9.3 データ`!loans_full_schema` → このデータは第 1 章と同一.

9.4 データ`!mariokart` → このデータは第 8 章と同一.

9.5 データ`!resume` → Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873 ここで示した分析のためのデータの幾つかの構造は議論しなかった. 実験計画ではブロッキング (blocking) を含むが, 典型的には 4 つの履歴書、各人種, 性別（名前で推測できる）, が送られている. このブロック化の影響は気にしなかったが、それは考慮したとしてもこの節の分析で得られた人種と性別の推定値はほとんど変わらず、標準誤差が小さくなっただけだからである. すなわち, もっとも重要な結論はより精緻な分析でも影響されることはない.