データ分析のための統計学入門

OpenIntro Statistics

Fourth Edition

(原著第 4 版, 翻訳初版第 2 刷)

(著者)

David M Diez [1)]

Mine Çetinkaya-Rundel [2)]

Christopher D Barr [3)]

(訳者)

国友直人[4)]

小暮厚之[5)]

吉田靖[6)]

[1)]データサイエンティスト (OpenIntro)

[2)]デューク大学准教授 (Duke University)

[3)]データサイエンティスト (Investment Analyst, Veradoro Capital)

[4)]統計数理研究所特任教授 (Institute of Statistical Mathematics), 東京大学名誉教授

[5)]東京経済大学教授 (Tokyo Keizai University)

[6)]東京経済大学教授 (Tokyo Keizai University)

目次

訳者 まえがき

本書は大学に入学して初めて統計学を学ぶ学生, 大学に進学を目指す高校生, ビジネスなどの諸分野でデータ分析をしている社会人のために書かれた書籍である. 2020 年春になり日本をはじめとして世界中の高等教育は「新型コロナ・ウイルス」のために深刻な危機に陥った. 日本の大学でもオンライン授業が開始されたが, 学生にとりよりどころとなる教科書がオンラインで利用可能であることが非常に少なかった. 訳者は私立大学文系における統計分野の授業のために必要に迫られ, 統計学分野についてかなり探したが内容的に十分と判断した教科書は見つからず, 学生諸氏に自由に利用してもらうことができなかった.

こうした中で幸いにも米国では既に OpenIntro なる NPO が配布している OpenIntro Statistics (4th Edition) という統計学分野への入門的な教科書があること分かった. 当初は無料で pdf ファイルをダウンロード可能（印刷物は有料）という手軽さに注目したが, 一読してみると実データを含め非常に内容が充実していること, (既に中級の教科書を出版している翻訳者代表にとっても) 日本で流布している多くの大学初級向けの教科書よりもむしろより適切であるように感じられたのである. そこで OpenIntro Statistics の著者にメールで連絡したところ, CCl(Creative Commons license) に違反しない限り自由に活用してよいとのことであった. 米国では大学で利用する教科書などの出版物の値段が高騰する中で一石を投じている高等教育における新しい流れのようである.

本書の著者は OpenIntro という NPO に属する有志のデータサイエンティスト, 大学教員, 投資ファンドのデータ分析家, である. 残念ながら 2021 年 1 月時点において日本語では本書のようなネット上で配布されている本格的な統計学, データ科学分野での教科書は利用可能ではないようである. 本書 (日本語版) の出版を 1 つの契機として, 日本においても関係するデータ分析や統計学の高等教育に関する議論が活発になることを希望する次第である. なお, 幸いにも NPO の日本統計協会のご厚意により印刷物を書籍の形 (有償・プリント版) で利用可能となった. 教科書として利用するときの補助教材 (原著・訳書に関する誤植・コメント, データ, その他) は日本統計協会のホームページ (https://www.jstat.or.jp/openstatistics/) からダウンロード可能である. なお, 本書は日本における統計学教育の標準的な認定制度となっている統計検定 3 級・2 級 (https://www.toukei-kentei.jp) の内容にほぼ準拠している.

また本書は入門という性格上で内容の基礎や計算の説明が不十分と感じる諸氏には例えば「統計学」(久保川達也・国友直人, 東京大学出版会),「R による統計データ分析入門」(小暮厚之, 朝倉書店),「(応用をめざす) 数理統計学」(国友直人, 朝倉書店), 英語なら原著 などの一読を薦めておこう.

日本語版の作成は, まず第 1 章〜第 3 章, 第 6 章, 第 8 章〜第 9 章を国友直人, 第 4 章〜第 5 章を吉田靖, 第 7 章を小暮厚之がそれぞれ担当, 内容に齟齬が生じないように調整を行って最終稿を作成した. 原著の誤植などは著者の了承のもとに修正したことを付け加えておく.

東京　　国友直人

2022 年 3 月

著者 まえがき

本書「データ分析のための統計学入門 (OpenIntro Statistics)」は社会における最近の課題を踏まえた統計学の学習の第一歩, 統計学のデータ分析への応用についての適切な入門, 明快かつ簡潔かつ学びやすさを目指している. 本書は主に大学生を念頭に置いて書かれているが, 場合によっては高校生, あるいは大学院生などにも適切な内容だろう. 著者は読者が統計的見方や方法の基礎を本書により理解すると共に, 次の 3 つの論点を理解してもらえることを希望している.

- 統計学は実際に幅広く利用されている応用分野である.
- 関心のある実際のデータを使って学ぶためには必ずしも数学の深い理解が必要というわけではない.
- 実際のデータは複雑であり, 統計学も完全ではない. しかし, 統計的分析の強みと弱みを理解することにより, 様々な世界を学ぶことに役立つ.

本書の概略

本書の各章はおおよそ次のような内容である.

1. データ分析への誘い. データの構造, 変数, および基本的なデータ収集の方法.

2. データの記述. データの要約法, グラフ, ランダム性を用いた推測の必要性.

3. 確率. 確率の基本的原理. 本章は後の章に必ず必要というわけではない.

4. 確率分布. 正規分布モデルと他の鍵となる確率分布.

5. 統計的推測の基本. 点推定, 信頼区間, 仮説検定についての入門. 母集団の比推定を利用した統計的推測の一般的な考え方.

6. カテゴリカル・データの推測. 正規分布やカイ二乗分布を用いた比率や分割表の推測.

7. 数値データの推測. t 分布を用いた 1 標本平均・2 標本平均の統計的推測, さらに 2 つあるいは多数の群データ比較のための分散分析 (ANOVA) 法.

8. 線形回帰への入門. 1 変数を用いた数値変数の回帰. この章の大部分は第 2 章の後に扱うことが可能.

9. 重回帰とロジット回帰. 複数の予測変数を用いた数値データやカテゴリカル・データの回帰分析.

学習目的により本書の内容を自由に選択したり順序付けることは可能である. 例えば主要な目的が第 9 章の重回帰分析をなるべく早く到達することなら, 次のように進むと良いだろう.

- 本書を通じてデータの構造とデータの記述法の入門を要約した第 1 章, 2.1 節, および 2.2 節.
- 正規分布の正確な理解のための 4.1 節.
- 統計的推測のコアを説明した第 5 章.

- t 分布の基本についての 7.1 節.
- 説明変数が 1 つの単回帰の考え方と方法について第 8 章.

例題と練習問題

例題 (Examples) は統計的方法がどのように応用されるか理解を深めるために用意した.

例題 0.1

一例を挙げる. ここでの質問への解答はどこにあるだろうか？

解答はここであり, 例の解答セクションにある.

読者が例への解答に用意ができているはずと判断する場合には確認問題 (Guided Practice) に述べる.

確認問題 0.2

読者は確認問題への解答を確かめるために脚注[1] にある解答により学ぶことができる. 読者はこの実際的な問題を解くことを強く勧める.

練習問題は各節の最後, 章末練習問題は各章の最後に与えてある. 奇数番の練習問題の解は付録 A に掲載されている.

補助教材

ビデオ収録教材, スライド, 統計ソフトラボ, 本書で利用したデータ, その他の教材は次の Web ページから利用可能である.

openintro.org/os

さらに第 4 版では本書で利用したデータについては付録 B に解説を加えた [**訳注**: データについては本書 Web ページにある.]. 各データ・セットについてはオンライン解説を以下の場所から利用できる: **openintro.org/data** および companion R package. Web(ウェブ) 情報を通じて誤植を含めコメントを歓迎している. 新たに見つけた誤植, これまでに分かっている誤植については次の Web(ウェブ Web 情報を参照 : **openintro.org/os/typos**. 高校レベルでの統計学に関心のある人に *Advanced High School Statistics* を用意したが, これは Leah Dorazio が本書の内容を高校の授業および AP 試験 ® [**訳注**: 米国の AP 試験は高校・大学初級の認定試験, AP-Statistics は日本の統計検定 3 級・2 級 https://www.toukei-kentei.jp のような役割を果たしている.] のために用意したものである.

謝辞

本プロジェクトは著者リストを超えて多くの熱心な関係者の貢献なしには実現できなかっただろう. 著者は OpenIntro のスタッフに感謝したい. さらに 2009 年に本書を掲示して以来, 価値あるフィードバックを提供してくれた多くの学生や教員にも感謝する. また本書をレビューしてくれた教員, Laura Acion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Juan Gomez, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, Steve Phelps, David Rockoff の諸氏に感謝する. これらの人々からのフィードバックにより様々な形で本書の内容を改善することができた.

[1] 確認問題 (Guided Practice) は思考の柔軟性を養うためのもので, 脚注にある解答により内容を確認できる.

第 1 章

データ分析への誘い

科学者は厳密な方法および注意深い観察値をもとにして疑問に答えようとしている. ここで観測値とはフィールド調査メモ, 調査, 実験などから集められるが, 統計的データ分析の鍵, **データ (Data)** と呼ばれる. 統計学はどのようにデータを集め, データ分析を行い, 結論が得られるか, 研究する分野である. 第 1 章ではまずデータの性質とデータの収集に焦点をあてよう.

日本語版の参考資料は **https://www.jstat.or.jp/openstatistics/** (日本統計協会) を訪問されたい.
原著の資料は以下にある. **www.openintro.org/os**

1.1 事例研究：ステントにより発作を抑える?

1.1 節では統計学における古典的な課題, 医療での治療効果の評価を例として説明しよう. この節およびこの章でよく利用する用語は後にもたびたび登場するが, 本節の目的は統計的分析が実際問題の解決に大きな役割を演じることを理解してもらうためである.

この節では心臓発作のリスクを抑えるために患者に行われているステント処置の脳梗塞の発作 (stroke) への効果を検討した実験データを考察する. ステント (stents) は血管内に挿入して心臓発作からの回復を助け, 更なる発作や死亡するリスクを抑制するための人工的装置である. 多くの医師は発作のリスクがある患者に治療効果があると期待していた. そこで何人かの医師により治療効果があるとの期待のもとに研究が行われた.

ステント (stents) 利用により発作リスクを下げられるだろう？

この課題に答えるために行われたある研究では 451 名のリスクを持つ患者を対象に実験が行われたが, ボランティアの患者はランダムに次の 2 つのグループに振り分けられた.

処理群 (treatment group). 処理群の患者はステント治療および医療サービスを受けた. この医療サービスには薬, リスク要因の管理, ライフスタイルの管理などが含まれていた.

対照群 (control group). 対照群の患者は処理群の患者と同様の医療サービスを受けたが, ステント治療は受けなかった.

このように研究者は 224 名の患者を処理群, 227 名を対照群に振り分けたが, この研究では対照群が処理群におけるステント治療の医学的インパクトを測ることができる比較点を提供してくれた. ステント治療の効果は 2 時点, 開始から 30 日間と 365 日に検証された. 例えばその中の 5 名の患者については図表 1.1 に要約されている. 患者の状態は脳梗塞の発作 (stroke), 発作なし (no event) に分類, 一定期間内で発作が起きたか否かを表している.

患者	群 (group)	0-30 日	0-365 日
1	処理 (treatment)	発作なし	発作なし
2	処理 (treatment)	発作	発作
3	処理 (treatment)	発作なし	発作なし
⋮	⋮	⋮	
450	対照 (control)	発作なし	発作なし
451	対照 (control)	発作なし	発作なし

図表 1.1: ステント研究の患者 5 名の例示

各患者から得られる個別のデータを調べることから最初の課題に対する回答を得るのは長く面倒な作業である. ここで統計的なデータ分析を行うことにより, すべてのデータを同時に考察することができる. 図表 1.2 はより理解可能な形でデータを要約したものである. 例えば処理群において 30 日以内に発作を経験した患者の数は処理群・発作の 33 を見ればよい.

	0-30 日		0-365 日	
	発作	発作なし	発作	発作なし
処理群	33	191	45	179
対照群	13	214	28	199
合計	46	405	73	378

図表 1.2: ステント研究の記述統計.

確認問題 1.1

確 処理群の 224 人の内最初の 1 年以内に 45 名の患者が発作を経験した. この数字から処理群の患者の中で 1 年以内に発作を起こした割合を計算しなさい[1]. (**注意**：本文中の確認問題への解答は注に与えられている.)

図表から統計量が求められるが, 要約統計量 (summary statistic) は多くのデータを要約している. 例えば研究での 1 年後の患者についての主要な結果は 2 つの数値, 処理群と対照群における発作経験の割合を見ればよいだろう.

処理群 (ステント) 内で発作を経験した患者の割合: $45/224 = 0.20 = 20\%$.

対照群内で発作を経験した割合: $28/227 = 0.12 = 12\%$.

この数値は群間の差を理解する上で有用であり, 処理群の患者 8%は追加的に発作を経験している！これには 2 つの重要な意味がある. 第 1 に医師が期待していたこと, ステントは発作の可能性を**少なくするだろう**という予見に反している. 第 2 に, 統計的な疑問, このデータは 2 つの群間の**本当の差**を示していると言えるか, である.

2 番目の疑問は微妙と言えるだろう. 例えばコイン投げを 100 回行ってみよう. コイン面が表の可能性が 50%であっても, 多分だが表が正確に 50 回だけは観察されないだろう. こうした変動はデータを発生させているほとんどすべての場合にあてはまる. すなわちステント研究における 8%の差はこうした自然的な変動であるかもしれない. しかしながら, データ数に対して観察される差が大きければ大きいほど, その差が単なる偶然によるとは信じられなくなる. そこで次のような疑問が生じる：この差は大きいので偶然生じたという考えを棄却すべきだろうか.

この疑問に正確に答える統計的準備はまだできていないが, 公表された研究の結論：この患者の研究はステント治療には危険性がある証拠となることを理解はできるだろう.

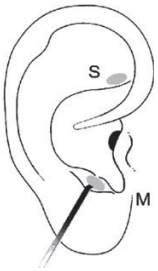
(**注意事項**) この研究の結果を患者全体についてのステント治療の全体に一般化してはならないことに注意しておく. この研究は自主的に参加した特定の患者についての結果であり, すべての患者についての結果というわけではない. また医療現場では様々な種類のステントが使われているが, ここでの研究は自己拡張型 (self-expanding)Wingspan ステント (Boston Scientific 社) という特定の機器についてのものである. ただしこの研究は重要な教訓, すなわち予断を持たずに結果に臨むべきことを示唆している.

[1] 患者 224 人の中で 365 日内に発作が起きた患者の割合：$45/224 = 0.20$ となる. [**訳注**：本書ではしばしば等号を四捨五入された数値でも用いる.]

練習問題

1.1 片頭痛と針治療, パート I. 片頭痛に対する針治療の効果を測定するあるランダム化実験が行われた. この統計的実験では 89 名の女性患者をランダムに処理群と対照群に振り分け, 43 名を処理群として針治療が行われ, 対照群の 46 名を対照群としてプラセボ (偽薬, この場合は間違った場所への針治療) が用いられた. 針治療の 24 時間後に片頭痛から解放されたか否かを問診されたが, 結果は次表に要約されている[2] .

		痛みの除去		
		はい	いいえ	全体
患者	処理群	10	33	43
	対照群	2	44	46
	合計	12	77	89



原論文から転載した図は片頭痛に対する適切な針治療の位置 (M) および不適切な針治療の位置 (S) を示している.

(a) 処理群の患者で針治療を受けた後, 24 時間に痛みから解放されたパーセントを求めよ.

(b) 対照群の患者で痛みから解放されたパーセントを求めよ.

(c) 針治療を受けて 24 時間後にどちらの群の患者がより痛みから解放されただろうか.

(d) この結果は片頭痛に苦しんでいるすべての患者にとり針治療が効果的であることを示唆しているだろうか. なおこの結論は観察事実に基づいた唯一の結論とは限らないだろう. 2 つの群に分けた患者への針治療による 24 時間後の効果の差についてどのような解釈が可能だろうか.

1.2 副鼻腔炎と抗生物質, パート I. 急性副鼻腔炎 (ふくびくうえん) に対する抗生物質治療として対象薬治療 (symptomatic treatments) の比較実験が行われた. 166 人の大人に対しランダムに 10 日間のアモキシリン (amoxicillin) の投与、見た目や味が抗生物質に似ているプラセボ投与により患者を処理群と対照群のどちらかにランダムに割り付けて調べた. プラセボ群には薬アセトアミノフォン (acetaminophen) や鼻炎薬 (nasal decongestants) などが用いられた. 10 日後に患者は症状の改善が見られたか否か問診されたが, 患者の反応は以下のようにまとめられている[3].

		改善 (自己申告) 症状		
		はい	いいえ	全体
患者	処理群	66	19	85
	対照群	65	16	81
	合計	131	35	166

(a) 処理群の患者で症状が改善したパーセントを求めよ.

(b) 対照群の患者で症状が改善したパーセントを求めよ.

(c) どちらの群の患者がより多く改善が見られただろうか.

(d) ここでの観察結果は鼻炎の症状を改善するために抗生物質とプラセボの効果の真の差を表しているとすると, その結論は観察結果から導かれる唯一の結論とは限らない. 鼻炎の症状を改善した抗生物質による処理群と対照群の患者の改善した割合についての観察された差を説明できる可能な他の案があるだろうか.

[2] G. Allais et al. “Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints”. In: *Neurological Sci.* 32.1 (2011), pp. 173–175.

[3] J.M. Garbutt et al. “Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial”. In: *JAMA: The Journal of the American Medical Association* 307.7 (2012), pp. 685–692.

1.2 データの形式

データを効果的に要約, 記述することは多くのデータ分析の第一歩である. この節では**データ行列** (data matrix) および本書で扱われるデータの要約, 様々な形式のデータ記述の基本的方法を導入する.

1.2.1 観測値・変数・データ行列

図表 1.3 はある融資仲介サービス (peer-to-peer lending, ソーシャル・レンディング, 一種の貸付組合) により提供されたデータから, ランダムに抽出された 50 ケースのデータをそれぞれ 1, 2, 3,...,50 行目に示している. 図表の各行はある資金貸付 (ローン) を示している. 各行の名前はある事例 (ケース), **観測単位**である. 各列は各ローンについての**変数 (variables)** と呼ばれる特性値を表している. 例えば第 1 行は 7,500 ドルのローン, 金利は 7.34%, 借り手の所在はメリーランド州 (Maryland, MD), 個人所得は 70,000 ドルである.

確認問題 1.2

確 図表 1.3 の第 1 行の等級 (grade) とは何だろうか, また最初の借り手の住居の所有状態は何だろうか? こうした疑問についての解答は脚注に与えてあるので確認してみよう[4].

ここでデータが含む重要な側面を明らかにするためには疑問を抱くことが重要である. 例えば各変数が何を意味しているか, 計測の単位は何か, 確認しておくことが必要である. ここでは図表 1.4 により変数の内容を説明しておこう.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	借家
2	25000	9.43	60	B	OH	254000	モーゲージ
3	14500	6.08	36	A	MO	80000	モーゲージ
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	借家

図表 1.3: データ loan50 のデータ行列から 5 列.

変数 (variable)	説明
loan_amount	貸付金額 (ローン金額,US ドル) .
interest_rate	ローンの金利 (年率) .
term	ローンの期間 (月数で計算) .
grade	ローンの等級, A-G のどれかに分類, ローンの質と返済の可能性を表す.
state	借り手が居住する US 州.
total_income	借り手の総所得, 副業からの所得を含み US ドルで表示.
homeownership	自宅, モーゲージ (抵当権付きの自宅), あるいは借家を示す指標.

図表 1.4: データ loan50 における変数と内容.

図表 1.3 のデータ形式は**データ行列** (data matrix) と呼ばれているが, スプレッド・シートで収集されているときにはデータを表現する便利で一般的な方法である. データ行列の各行は事例 (観測単位), 各列は各変数の観測値に対応している.
データを記録する場合には他の形式を使いたい理由がない限り, データ行列を使うべきである. この形式なら新しい観測値は 1 行を加える, 新たな変数は 1 列を加えれるだけでよい.

[4] ローンの等級は A, 借り手の住居は借家である.

確認問題 1.3

確 授業の課題, クイズ, 試験などの成績はしばしばデータ行列の成績表に記録される. それではどのようにデータ行列を使って成績をつけるだろうか[5].

確認問題 1.4

確 米国内 3,142 の郡 (counties) データを考察するが, 郡の名前, 属する州, 2017 年の人口, 2010 年-2017 年の人口変化, 貧困率, その他 6 個の特性値が得られる. ではデータ行列はどう作れるだろうか[6].

確認問題 1.4 で説明しているデータは郡データを表し, 図表 1.5 のデータ行列を示している. 変数は図表 1.6 で説明されている.

[5] 例えば次のような方法がある. 各学生に各行を割り当て, 課題, クイズ, 試験を行う度に列を付け加えていく. この方法ならある学生の評価履歴が一行で分かる. また各列に学生の名前などの情報をつけることも容易となる.

[6] 各郡を 1 ケースと見て 11 の属性情報を記入する. 3,142 行 11 列の表によりデータが表現され, 各行には郡, 各列には郡の情報が示される.

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	hs_diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	yes	hs_diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	hs_diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	hs_diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	hs_diploma	40041
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

図表 1.5: データ county からの抜粋

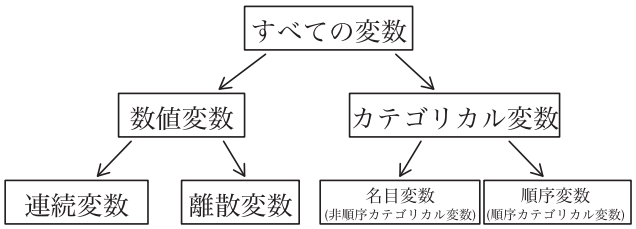
変数	説明
name	郡名 (county name) .
state	郡が属する州またはコロンビア特別区.
pop	2017 年の人口.
pop_change	2010 年から 2017 年にかけての人口の変化率. 例えば最初の列 1.48 は 2010 年から 2017 年にかけて人口が 1.48% 増加を示している.
poverty	人口に占める貧困層の割合.
homeownership	人口に占める自宅所有あるいは同居者 (持ち家の親と同居の子供) .
multi_unit	多重構造住居, アパートに居住している割合.
unemp_rate	失業率.
metro	郡が都市部を含むか否か.
median_edu	教育の中位レベル (高校以下, 高卒, 大学, 大卒: それぞれ below_hs, hs_diploma, some_college, bachelors).
median_hh_income	郡における家計所得の中位数. ただし家計における 15 歳以上の構成員の総所得.

図表 1.6: データ county の変数と説明.

1.2.2 変数のタイプ

ここでデータ county における失業率, 人口, 州, 中位教育水準を見てみよう. これらの変数はそれぞれ他とは異なる特徴があるが, 幾つかの共通の特質も備えている.

第 1 に, 失業率 (変数 unemp_rate) は様々な値を取り得る数値変数 (numerical) と呼ばれるが, 互いに足したり, 引いたり, 平均を取ることが意味がある変数である. 他方, 電話番号コードは数値とは呼ばないが, これは幾つかのコード番号を足したり, 和をとったり, 差をとることの意味がはっきりしないからである. 変数 pop は数値変数あるが, 失業率とは少し意味が異なる. 人口を表すこの変数は非負整数 0,1,2,... の値をとる. このことから人口変数は**離散的** (discrete) と呼ばれるが, 数値としては実数ではなくとびとびの値をとる, これに対して失業率は**連続的** (continuous) と呼ばれる. 変数 state はワシントン DC を除けば 51 個の値をとり, アラバマ州 (AL), アーカンソー州 (AK), ワイオミング州 (WY) などである. ここでの変数は幾つかの分類 (カテゴリー) の値をとるので変数 state はカテゴリカル (categorical, 非数値, 質的) 変数と呼び, 取りうる可能な値を水準 (levels) と言う. 最後に教育中位水準を取り上げると, これは郡の居住者の中位教育水準を示し, 高校以下 (below-hs), 高卒 (hs-diploma), 大学進学 (some-college), 大卒 (bachelors) などの値をとる. この変数は一種のハイブリッド, カテゴリカル変数ではあるが順序付けられている. こうした性質を持つ変数は順序的 (ordinal) と呼ばれ, 順序に意味を持たないカテゴリカル変数を名目的 (nominal) と呼ぶ. データ分析を簡単化するために本書では名目的変数は名目的 (順序付けられない) カテゴリカル変数とする.



図表 1.7: 変数タイプの分類.

例題 1.5

ある統計学の講義を聴講している学生のデータである. 各学生について兄弟・姉妹, 身長, 既に統計学のコースを履修したかどうか, という変数の値が記録されている. この変数を連続型数値, 離散型数値, カテゴリカル型に分類しなさい.

例

兄弟・姉妹の身長は数値変数である. 兄弟・姉妹の数は正整数なので離散型である. 身長は連続的に変化するので連続型である. 最後の変数は 2 つのカテゴリー値をとる変数, 既に統計学の講義を履修したか否かであるから, カテゴリカル変数である.

確認問題 1.6

片頭痛 (migraines) への新しい治療薬の効果を測定する実験を考えよう. 実験変数として各患者のグループ分け, 処理群と対照群を用いる. 変数 num-migraines は 3 か月に患者が経験する片頭痛の回数を表すとする. この変数は数値変数, あるいはカテゴリカル変数だろうか[7].

確

1.2.3 変数間の関係

多くの研究者は 2 つ, あるいはそれ以上の変数間の関係を求めて研究を行っている. 例えばある社会科学者は次のような疑問に対する答えを探しているとしよう.

(1) ある郡の住宅保有が全国平均より低ければ, その郡の多層住宅 (multi-unit structures) の居住率は全国平均より高い, あるいは低い傾向があるだろうか.

(2) 人口増加が平均より多い郡の中位家計所得は他の郡より高い (あるいは低い) 傾向があるだろうか.

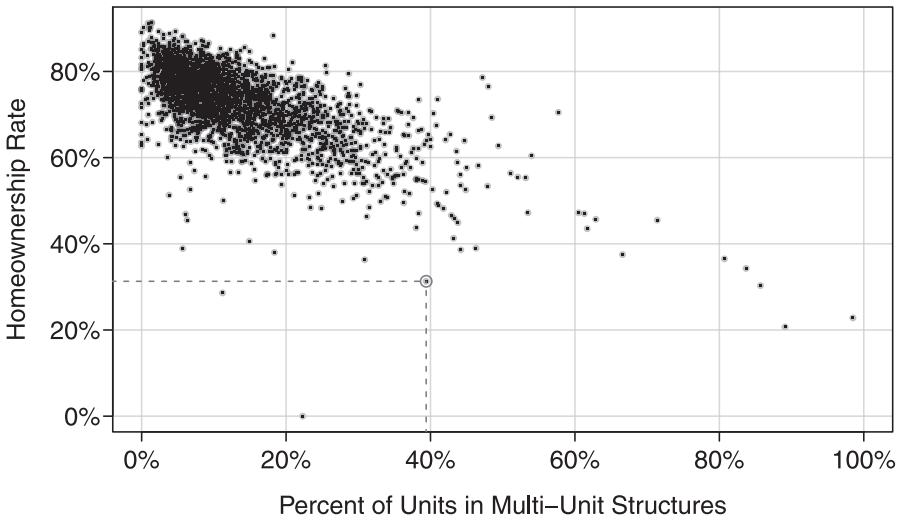
(3) 米国の郡データでは家計の中位所得は中位の教育水準の予測量として役に立つだろうか.

このような疑問に答えるために図表 1.5 で示すようなデータ *county* を集める必要がある. このデータの要約統計量は郡に関する 3 つの疑問へ鍵を与えてくれる. またグラフはデータを視覚的に表現してくれる. 散布図 (scatterplot) はグラフの一種であるが 2 つの数値変数の関係を研究するために用いられる. 図表 1.8 は自宅保有とアパート居住・多層構造 (つまりアパートやコンドミニアム) 居住を比較したものである. 図表の中の各点は各郡を示している. 例えばハイライトされている 1 点は郡データの郡 413 ：ジョージア州のチャタフーチ (Chattahoochee) 郡では多層構造 39.4%, 自宅所有 31.3%である. この散布図は 2 つの変数の間の関係, 多層構造住居が多ければ自宅所有は低いことを示している. この関係はなぜなのか, その理由を挙げてそれぞれ最も妥当な説明を考察してみると良いだろう.

この散布図には明確なパターンがあるので多層構造住居率と自宅保有率は関連している. 2 つの変数が互いに関連しているとき, 変数を関連 (associated) 変数と呼ぼう. 関連する変数は互いに従属 (dependent) 変数とも呼ばれる. [**訳注**：日本語ではしばしば相関があると呼ばれる.]

---

[7] 実験グループを示す変数は 2 値の内 1 つをとるのでカテゴリカル変数である. 変数 num-migraines は片頭痛の回数を示しているので離散数値変数である.



図表 1.8: 郡における自宅保有と多層構造住居率の散布図 (点はジョージア州のチャタフーチ郡, 多層率 39.4%と自宅保有率 31.3%, を示す).

確認問題 1.7

確 図表 1.4 で説明されているデータ *loan50* の変数を調べてみよう. このデータの中で関心のありそうな変数間における関係について課題を設定してみよう[8].

例題 1.8

図表 1.9 の散布図にある郡の人口の 2010 年から 2017 年への変化と家計所得の中位数 (中央値,Median) との関係を調べると, これらの変数に関係があるだろうか.

例

郡の家計所得の中位数が高ければ郡で観察される人口成長が高い. この関係はすべての郡について正しいとは言えないが, こうした傾向にあるのは図表から確かである. したがってこれらの変数には何らかの関係があり変数は関連 (associated) している.

図表 1.8 には負のトレンドがあり, 多層構造住居が多ければ自宅保有は少ない傾向, 負の関係 (negative association) がある. 正の関係 (positive association) の例は図表 1.9 に見られる中位所得と人口変化, 中位所得が多ければ人口増加が大きい, などである. 2 つの変数に関連がなければ独立 (independent) と呼ばれる. つまり 2 つの変数間に明らかな関係を見いだせなければ独立と呼ばれる.

関連しているか, あるいは独立か, 並立はあり得ない

2 つの変数は何らかの意味で関連している (associated) 場合は独立ではなく, 同時に関連してかつ独立 となることはない.

[8] 2 つの疑問 (1) ローン金額と総所得の関係はないか. (2) 平均所得より上であれば金利はより高い, あるいは低いか.