# 『よくわかるデータリテラシー　データサイエンスの基本』
# インストラクションガイド

## 初版第１刷への訂正

１．p.76、11 行

（誤）標本の数値は正規分布に近づき、

（正）標本平均の分布は正規分布に近づき、

２．p.103、下から 4 行

（誤）１－見逃し率

（正）１－誤検出率

３．p.128、第 4 講(4)

（誤）極微の世界から宇宙の果てまで、長さを 10 倍ずつしながら

（正）宇宙の果てから極微の世界まで、長さを 1/10 ずつしながら

## 「付録－－さらに勉強したいときは」への本の追加

全般

(18) ダニエル・カーネマン : ファスト＆スロー　あなたの意思はどのように決まるか？（上）（下）、早川書房、2012

著者は、認知科学を経済学に導入して行動経済学を確立した功績でノーベル経済学賞を受けています。豊富な話題についてわかりやすく解説した素晴らしい本です。

(19) カール・T.・バーグストローム、シェヴィン・D.・ウエスト : デタラメ　データ社会の嘘を見抜く、日本経済新聞出版、2021

(20) ピーター・シュライバー：統計データの落とし穴～その数字は真実を語るのか？～、ニュートンプレス、2021

原題の”Bad Data”も、訳書のタイトルも内容を適切に表していません。不適切な評価指標や、評価指標の不適切な運用の例をいっぱい挙げています。

**授業計画へのヒント**

この本は 15 講からなっていますが、90 分 15 回の授業には内容が足りないと思います。どんな内容を追加するかについては、次の 2 つのモデルカリキュラムが参考になります。

一つは、一般教育（共通教育）におけるデータリテラシーのモデルカリキュラムです。2020 年 4 月に、数理・データサイエンス教育強化拠点コンソーシアムから

「数理・データサイエンス・AI（リテラシーレベル）モデルカリキュラム ～　データ思考の涵養　～」

http://www.mi.u-tokyo.ac.jp/consortium/pdf/model_literacy.pdf

が出されています。その 12 ページに、図 1 に示す「データリテラシー＜スキルセット＞」があります。

図中で✔を付けた項目は、この本で扱っています（一部、取り上げていない話題があります）。

**2. データリテラシー＜スキルセット＞**

| 2.データリテラシー | キーワード（知識・スキル） |
|---|---|
| 2-1. データを読む | ✓・データの種類（量的変数、質的変数）<br>✓・データの分布(ヒストグラム)と代表値（平均値、中央値、最頻値）<br>✓・代表値の性質の違い（実社会では平均値＝最頻値でないことが多い）<br>✓・データのばらつき（分散、標準偏差、偏差値）<br>✓・観測データに含まれる誤差の扱い<br>・打ち切りや脱落を含むデータ、層別の必要なデータ<br>✓・相関と因果（相関係数、擬似相関、交絡）<br>✓・母集団と標本抽出（国勢調査、アンケート調査、全数調査、単純無作為抽出、層別抽出、多段抽出）<br>✓・クロス集計表、分割表、相関係数行列、散布図行列<br>✓・統計情報の正しい理解（誇張表現に惑わされない） |
| 2-2. データを説明する | ✓・データ表現（棒グラフ、折線グラフ、散布図、ヒートマップ）<br>・データの図表表現（チャート化）<br>・データの比較（条件をそろえた比較、処理の前後での比較、A/Bテスト）<br>✓・不適切なグラフ表現（チャートジャンク、不必要な視覚的要素）<br>・優れた可視化事例の紹介（可視化することによって新たな気づきがあった事例など） |
| 2-3. データを扱う | ✓・データの集計（和、平均）<br>・データの並び替え、ランキング<br>・データ解析ツール（スプレッドシート）<br>・表形式のデータ（csv） |

図 1　データリテラシー＜スキルセット＞

もう一つは、専門教育における数理統計学のカリキュラム標準です。2021 年 4 月に情報処理学会から

「データサイエンス・カリキュラム標準（専門教育レベル）」

https://www.ipsj.or.jp/annai/committee/education/public_comment/qe83kf0000002hlu-att/a1618203503118.pdf

が発表されています。

その 2～4 ページにある「数理統計学」の部分を図 2 に示します。図中で✔を付けた項目は、この本で扱っています（一部、取り上げていない話題があります）。

## 3. 数理統計学

種別 知識

| | 通し番号 | 優先度 | 割り当て時間数(h) | DS-BoK KA番号 |
|---|---|---|---|---|
| | DS-001a | T1 | 0.50 | KA01.01 |
| | 場合の数，順列・組みわせの概念を理解している． | | | |
| | 【備考】高校数学Aの範囲（履修していない高校生も存在） | | | |
| | DS-002 | T1 | 0.00 | KA01.01 |
| | 確率の概念を理解し，同時確率と条件付き確率の意味や違いを説明できる． | | | |
| | 【備考】数理・データサイエンス・AIモデルカリキュラム（リテラシーレベル）によりカバー | | | |
| | DS-003 | T1 | 0.00 | KA01.01 |
| ✓ | 平均（相加平均），中央値，最頻値の算出方法の違いを説明できる． | | | |
| | 【備考】数理・データサイエンス・AIモデルカリキュラム（リテラシーレベル）によりカバー | | | |
| | DS-004a | T1 | 0.00 | KA01.01 |
| ✓ | 分散と標準偏差の意味と定義を説明できる． | | | |
| | 【備考】数理・データサイエンス・AIモデルカリキュラム（リテラシーレベル）によりカバー | | | |
| | DS-005 | T1 | 0.50 | KA01.01 |
| ✓ | 母（集団）平均と標本平均，不偏分散と標本分散の違いを説明できる． | | | |
| | DS-006 | T1 | 0.50 | KA01.01 |
| ✓ | 標準正規分布の分散と平均の値を知っている． | | | |
| | DS-007 | T1 | 0.00 | KA01.01 |
| ✓ | 相関関数と因果関係の違いを説明できる． | | | |
| | 【備考】数理・データサイエンス・AIモデルカリキュラム（リテラシーレベル）によりカバー | | | |
| | DS-008 | T1 | 0.25 | KA01.01 |
| ✓ | 名義尺度，順序尺度，間隔尺度，比例尺度の違いを説明できる． | | | |
| | DS-009 | T1 | 0.00 | KA01.01 |
| ✓ | 一般的な相関係数（ピアソン）の分母と分子を説明できる． | | | |
| | 【備考】数理・データサイエンス・AIモデルカリキュラム（リテラシーレベル）によりカバー | | | |
| | DS-010 | T1 | 1.00 | KA01.01 |
| | 5つ以上の代表的な確率分布を説明できる． | | | |
| | DS-011 | T1 | 0.50 | KA01.01 |
| ✓ | 二項分布の事象もサンプルサイズが増えていくとどのような分布に近似されるかを知っている． | | | |
| | DS-012a | T1 | 1.00 | KA01.01 |
| | ピアソンの積率相関係数とクラメールの連関係数とスピアマンの順位相関係数の適用場面の違いを知っている． | | | |
| | DS-013 | T1 | 1.00 | KA01.01 |
| | ベイズの定理を説明できる． | | | |

| | | | | | |
|---|---|---|---|---|---|
| | DS-028 | E | 1.00 | KA01.01 | KA01.05 |
| | 点推定と区間推定の違いを説明できる. | | | | |
| | DS-029 | E | 1.00 | KA01.01 | KA01.05 |
| ✓ | 帰無仮説と対立仮説の違いを説明できる. | | | | |
| | DS-030 | E | 1.00 | KA01.01 | KA01.05 |
| ✓ | 第1種の過誤，第2種の過誤，p値，有意水準の意味を説明できる. | | | | |
| | DS-031 | E | 1.00 | KA01.01 | KA01.05 |
| ✓ | 片側検定と両面検定の違いを説明できる. | | | | |
| | DS-032a | E | 2.00 | KA01.01 | KA01.05 |
| | データ間に対応のある場合と無い場合の検定手法の違いを説明できる. | | | | |
| | IPSJ-09 | E | 2.00 | KA01.01 | |
| | 平均値の検定と平均値の差の検定(群間の対応あり，なしを含むt検定)を理解している. | | | | |

種別 スキル

| | 通し番号 | 優先度 | 割り当て時間数(h) | DS-BoK KA番号 | |
|---|---|---|---|---|---|
| | DS-001b | T1 | 0.50 | KA01.01 | |
| | 順列や組合せを式 nPr，nCr を用いて計算できる. | | | | |
| | 【備考】高校数学Aの範囲(履修していない高校生も存在) | | | | |
| | DS-004b | T1 | 0.00 | KA01.01 | |
| ✓ | 与えられたデータにおける分散と標準偏差が計算できる. | | | | |
| | 【備考】数理・データサイエンス・AIモデルカリキュラム(リテラシーレベル)によりカバー | | | | |
| | DS-012b | T1 | 0.50 | KA01.01 | |
| | 変数が量的，質的どちらの場合でも関係の強さを算出できる. | | | | |
| | DS-032b | E | 2.00 | KA01.01 | KA01.05 |
| | 推定する対象となるデータの対応の有無を考慮した上で適切な検定手法を選択し，適用できる. | | | | |

図 2　数理統計学

図１、図２に示されている、この本でカバーしていない項目を適宜選んで付け加えて、15週の授業を組み立ててください。一つの例を次の表に示します。

| | |
|---|---|
| 第 1 回 | 講義概要、第 0,1 講 |
| 第 2 回 | 第 2 講,第 3 講の前半 |
| 第 3 回 | 第 3 講の後半,第 4 講 |
| 第 4 回 | 第 5 講 |
| 第 5 回 | 第 6 講 |
| 第 6 回 | 第 7 講 |
| 第 7 回 | 第 8,9 講 |
| 第 8 回 | 第 10 講 |
| 第 9 回 | 仮説検定の例と演習　たとえば、二項分布、平均の差 |
| 第 10 回 | 第 12,13 講（第 11 講は読ませるだけ） |
| 第 11 回 | 第 14,15 講 |
| 第 12 回 | 第 11 回の内容から選んで、ディベートあるいは少人数に分けて討論 |

他の 3 回は、次の内容から選んで、適当な回で講義および演習を行います。

・順列・組合せ

・確率、条件つき確率

・ベイズの定理、その応用

・回帰分析・重回帰分析

・統計ソフト R の使いかた

中間テストを行うか、小テストを何回か行います。

次の表に各講のページ数を示しましたので、参考にしてください。

| | | | |
|---|---|---|---|
| 第 0 講 | 4 | 第 9 講 | 7 |
| 第 1 講 | 11 | 第 10 講 | 7 |
| 第 2 講 | 8 | 第 11 講 | 8 |
| 第 3 講 | 6 | 第 12 講 | 11 |
| 第 4 講 | 7 | 第 13 講 | 6 |
| 第 5 講 | 8 | 第 14 講 | 5 |
| 第 6 講 | 11 | 第 15 講 | 9 |
| 第 7 講 | 10 | | |
| 第 8 講 | 9 | 計 | 127 |

## 演習の追加

p.22　［演習 2.8］2050 年の日本の GDP 世界ランキングの予想について、いくつかのウェブ・データを調べなさい。

p.34　［演習 4.6］Kg、MB、mℓのように、単位の前に K（キロ）、M（メガ）、m（ミリ）を付けて、千倍、百万倍、千分の一を表します。この役割をする主な記号は、$10^3$ や $10^{-3}$ おきになっています。次の表の空欄を埋めましょう。

| 意味 | 記号 | 読み | 意味 | 記号 | 読み |
|---|---|---|---|---|---|
| $10^3$（千） | K | キロ | $10^{-3}$（千分の一） | m | ミリ |
| $10^6$（百万） | M | メガ | $10^{-6}$（百万分の一） | | マイクロ |
| $10^9$（十億） | | ギガ | $10^{-9}$（十億分の一） | n | |
| $10^{12}$（兆） | T | | $10^{-12}$（一兆分の一） | p | |
| $10^{15}$（千兆） | P | ペタ | $10^{-15}$（千兆分の一） | f | フェムト |
| $10^{18}$（百京） | E | エクサ | $10^{-18}$（百京分の一） | a | アット |

記憶容量や回線容量の増大にともなって、T や P や E を目にする機会が増えました。小さいほうも、集積回路やウイルスの寸法、有害物質の濃度などで n, p, f が使われます。

このほか、百倍を表す h（ヘクト、ヘクタールやヘクトパスカル）、1/10 を表す d（デシ、デシリットル）、1/100 を表す c（センチ、センチメートル）が使われることもあります。

μ は長さを表すときは、ミクロンと呼ばれ、$10^{-6}$m を意味します。mμ はミリミクロン、$10^{-9}$m です。これらは、μm、nm と書くのが正式です。重量を表すには、Mg、Gg、Tg の代わりに、トン、キロトン、メガトンがふつう使われます。容積では、mℓを cc とも表記します。

p.63　［演習 7.6’］演習 7.6 が面倒なのは、プロ野球ですから、1 ヶ月の試合数が大きく違ってはおかしいからです。次の問題に代えれば容易になります。

ある大学には 2 つの学部があります。ある年、どちらの学部も女子のほうが合格率が高かったのに、大学全体では男子のほうが合格率が上でした。そういう合格者数の例を作ってください。

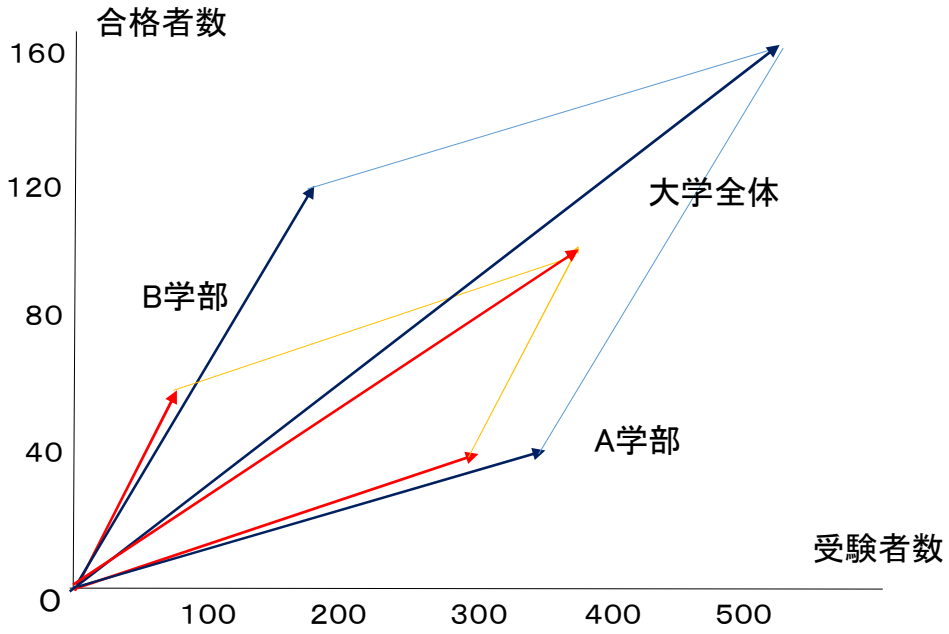
（解答例）次の表に示します。数字は、合格者数／受験者数（合格率）です。

| | A 学部 | B 学部 | 大学全体 |
|---|---|---|---|
| 男子 | 40／350（11.4%） | 120／175（68.6%） | 160／525（30.5%） |
| 女子 | 40／300（13.3%） | 60／75（80.0%） | 100／375（26.7%） |

このことは、次の図で考えるとわかりやすく理解できます。横軸は受験者数、縦軸は合格者数です。受験者数、合格者数を同時に示すと、原点からのベクトルになります。青は男子、

赤は女子のベクトルです。合格率は、合格者数／受験者数ですから、ベクトルの勾配で示されます。受験者数軸に近いほど、合格率が低いことになります。

Ａ学部、Ｂ学部の受験者数・合格者数にたいして、大学全体の受験者数・合格者数はベクトルの足し算、つまり図の平行四辺形の第４の頂点へのベクトルで表されます。そうすると、図のとおり、どちらの学部も女子のほうが合格率が高かったのに、大学全体では男子のほうが合格率が高くなっていることがわかります。




p. 120　［演習 15.1］あるテーマを選んで、一次情報と二次情報の両方を用いて調査しなさい。一次情報と二次情報の違いについて論じなさい。

この演習は、プレイディみかこの息子が、イギリスの公立中学で宿題として出されたそうです（プレイディみかこ・鴻上尚史：何とかならない時代の幸福論、p.199、朝日新聞出版、2021）。