はじめての統計学

道家　暎幸
伊藤　真吾
宮﨑　直
酒井　祐貴子
共著

コロナ社

ま え が き

本書は，初めて統計学を学ぶ人のための入門書として書かれたものである。

近年の情報化社会では，さまざまな分野においてデータ解析の方法論としての統計学の重要性が増してきている。大学においても理系・文系を問わず，多くの分野で統計学の科目が設けられ，それに伴い統計学に関する教科書，参考書が多数出版されている。その中には，統計の理論的な側面に重点を置き，初学者には理解し難いテキストや，単なるお話に終始し統計的な手法を会得できない内容のテキストもある。本書は実用面を意識し，データ解析のための基本的な手法の理解と，それを使って結果を解釈できる能力を養うことを目標に書かれた教科書である。したがって，統計理論の証明は極力避け，その理論を使うための説明と理解度を深めるための例題に重点を置いた。それに関連し，応用力を養うための演習問題も精選した。

本書は大学の通年科目（4 単位）での使用を想定しているため，統計学の入門書として基本的な内容を一通り網羅したつもりである。半期科目（2 単位）で使用される場合は，各章での主要な部分を抜粋し学習することも可能である。本書で学ぶ上で，高等学校までの教育課程における確率統計の予備知識は必要としないが，数学 II 程度の微分積分の知識があった方が望ましい。

本書は初めて統計学を学ぶ人が，統計学の内容を楽しく理解できるように，平易な文章で，多くの例題や図表を取り入れ，自学自習でも興味を持って無理なく読み進めることのできるように心掛けた。本書で学んだ方々が，各分野で統計的な手法を適用し，あるいはより高度な統計学に興味を持ち，新しい研究の動機付けとなれば，著者等の喜びとするところである。

本書を執筆するに当たり，著者間で遺漏のないように意見交換したつもりであるが，不備な点もあると思われる。統計教育に携わっておられる教授各位のご批判と，本書で学ばれた方々からのご意見をお寄せいただければ幸いである。

本書の刊行に際し，ご尽力いただいたコロナ社の方々に深甚なる謝意を表すものである。

2017 年 1 月

著 者 一 同

目　　次

1. データの整理

1.1 集団と変数の分類 …… 1
1.2 度数分布表とヒストグラム …… 2
1.3 代表値と散布度 …… 8
1.3.1 代　表　値 …… 8
1.3.2 散　布　度 …… 12
1.4 2次元データ …… 18
1.4.1 相関係数 …… 18
1.4.2 回帰直線 …… 22

2. 確　　率

2.1 集　　合 …… 26
2.2 順列と組合せ …… 28
2.2.1 順　　列 …… 28
2.2.2 組　合　せ …… 29
2.3 事象と確率 …… 32
2.3.1 試行と事象 …… 32
2.3.2 確率の定義 …… 34

2.3.3 確率の法則 …… 37
2.4 条件付き確率と乗法定理 …… 40
2.5 ベイズの定理 …… 43
2.6 反復試行の確率 …… 47

3. 確率分布

3.1 確率変数と確率分布 …… 49
3.2 離散型確率分布 …… 51
3.3 二項分布 …… 57
3.4 ポアソン分布 …… 60
3.5 いろいろな離散型確率分布 …… 64
3.5.1 離散型一様分布 …… 64
3.5.2 超幾何分布 …… 65
3.6 連続型確率分布 …… 66
3.7 正規分布 …… 71
3.7.1 標準正規分布 …… 73
3.7.2 確率変数の標準化 …… 77
3.7.3 正規分布による二項分布の近似 …… 78
3.8 いろいろな連続型確率分布 …… 82
3.8.1 連続型一様分布 …… 82
3.8.2 指数分布 …… 82

4. 標本分布

4.1 標本調査 …… 84
4.2 母集団分布と標本分布 …… 85

4.3　母比率と標本比率 …… 91
4.4　正規母集団の標本分布 …… 94
4.4.1　χ^2　分　布 …… 94
4.4.2　t　分　布 …… 97
4.4.3　F　分　布 …… 100

5. 推　　定

5.1　点　推　定 …… 104
5.1.1　点推定の考え方 …… 104
5.1.2　不偏性，有効性，一致性 …… 105
5.2　区　間　推　定 …… 109
5.2.1　区間推定の考え方 …… 109
5.2.2　母平均の区間推定 …… 110
5.2.3　母分散の区間推定 …… 118
5.2.4　母比率の区間推定 …… 120

6. 仮　説　検　定

6.1　仮説検定の考え方 …… 123
6.2　母平均の検定 …… 128
6.3　母分散の検定 …… 134
6.4　母比率の検定 …… 136
6.5　母平均の差の検定 …… 139
6.6　等分散の検定 …… 144
6.7　適合度の検定 …… 146
6.8　独立性の検定 …… 149

7. 分散分析法

7.1 分散分析法とは …… 155
7.2 1元配置法 …… 156
7.2.1 実験順序の無作為化 …… 157
7.2.2 平方和の分解 …… 158
7.2.3 検定方法 …… 160
7.3 2元配置法 …… 165
7.3.1 交互作用とは …… 165
7.3.2 実験順序の無作為化 …… 166
7.3.3 平方和の分解 …… 167
7.3.4 検定方法 …… 170

付録 …… 177
A.1 数表 …… 177
A.2 問題演習における数値計算上の注意 …… 189
引用・参考文献 …… 190
演習問題解答 …… 191
索引 …… 207

1 データの整理

統計学の内容は「記述統計」と「推測統計」の二つに大別される。この章では，調査，実験によって得られたデータから集団の性質や傾向を把握するための方法である「記述統計」を学ぶ。一般に，データは数値などの大量な情報の集まりなので，そのまま漠然と眺めていても集団が持つ性質や傾向は見えてこない。集めたデータの度数分布表からヒストグラムを作成して視覚的に，また平均や分散などを求めて数値的に集団の傾向を捉える方法を学習していこう。

1.1 集団と変数の分類

ある市の A 中学校の生徒の**特性**について調査を行うとする。調査の対象は A 中学校の生徒全体とし，調査項目は生徒の性別，身長，体重，世帯人数，習い事の数，学校生活の満足度などである。このような調査を実施して得られる統計資料を**データ**という。また，調査対象の全体を**集団**，調査される個々の対象を**個体**，調査される項目を**変数**という。

変数は，性別，学校生活の満足度のようにデータがカテゴリーで表される**定性的変数**（質的変数）と身長，体重，世帯人数，習い事の数のようにデータが数値（観測値，特性値）で表される**定量的変数**に大別される。定性的変数の調査に関して，性別は「男性」,「女性」の二つのカテゴリーからなる属性であり，このどちらかを選択させることで質的データを得る。また，学校生活の満足度は「満足」,「やや満足」,「どちらともいえない」,「やや不満」,「不満」のようにいくつかのカテゴリーを作成し，いずれかを選択させることで質的データを得

る。一方，定量的変数は，身長，体重のように，ある区間内のすべての実数の値をとり得る**連続型変数**と，世帯人数，習い事の数のように $0, 1, 2, \cdots$ といったとびとびの値をとる**離散型変数**に分類される。連続型変数を扱う際は，得られた数値について測定精度を考え，四捨五入して丸めたり，有効数字で表すことも必要である。例えば，身長のような連続型変数は，精密な測定機器を用いれば 164.72 cm のように詳しく測定することが可能であるが，この結果は測定時点によって変化すると考えられるし，一般的な身長調査を目的とするならば，それほど詳しい数値は意味がない。そこで，小数第 1 位を四捨五入して 165 cm と丸めれば十分である。

変数
- 定性的変数（質的変数）$\cdots$ 血液型，性別，好きな歌手 など
- 定量的変数（数値で表される）
  - ・連続型変数 $\cdots$ 身長，体重，時間，距離 など
  - ・離散型変数 $\cdots$ サイコロの目，人数，事故件数 など

1.2　度数分布表とヒストグラム

データから有効な情報を引き出すためには，データを目的に合わせて整理し，その特徴をわかりやすくまとめる必要がある。本節では，おもに定量的変数についてのデータのまとめ方を解説する。定量的変数を扱う場合，データの分布の中心やばらつきの大きさを知ることが重要であり，そのためには度数分布表とヒストグラムを作成することが有効である。

集めたデータがどのように分布しているかを表の形でまとめたものを**度数分布表**という。例えば，あるクラスの学生の身長の分布を調べたい場合，身長の範囲をいくつかの区間に区切り，各区間に属する学生数の分布を表にする。この区間のことを**階級**，各階級の中央の値を**階級値**，各階級に入るデータの個数を**度数**という。**表 1.1** はあるクラスの学生 50 人の身長〔cm〕のデータ，**表 1.2** は表 1.1 をもとに作成された度数分布表である。

表 1.1　あるクラスの身長のデータ

172	165	156	166	164	146	165
152	165	150	162	153	148	166
155	157	170	167	157	145	171
160	154	150	171	159	169	168
153	156	158	147	149	160	161
164	146	155	163	152	160	156
158	155	155	172	162	154	151
164						

表 1.2　あるクラスの身長の度数分布表

階級 以上　未満	階級値	度数
145 ～ 150	147.5	6
150 ～ 155	152.5	9
155 ～ 160	157.5	12
160 ～ 165	162.5	10
165 ～ 170	167.5	8
170 ～ 175	172.5	5
合　計	–	50

ここで，度数分布表の作成手順を紹介しておこう†。

(1)　データの最大値 M と最小値 m を見つける。

(2)　階級の数 k を決め，$a_0 = m$，$a_k = M$ とおく。

k の値は大きすぎると全体の傾向を捉えにくく，小さすぎると部分的な特性がわからない。明確な決まりはないが，データの個数 n に応じて

n が 30 前後のときは　$4 \leqq k \leqq 6$,

n が 50 前後のときは　$5 \leqq k \leqq 7$,

n が 100 前後のときは　$7 \leqq k \leqq 10$,

n が 100 以上のときは　$10 \leqq k \leqq 20$

を目安に設定するとよい。

(3)　a_0 から a_k の範囲を k 等分し，分点を小さい方から順に $a_1, a_2, \cdots, a_{k-1}$ とおく。

これにより，階級は $a_0 \sim a_1$，$a_1 \sim a_2$，$\cdots$，$a_{k-1} \sim a_k$ となる。各階級 $a_{i-1} \sim a_i$ $(i = 1, 2, \cdots, k)$ ごとに，a_{i-1} を級下限界，a_i を級上限界といい，これらをまとめて**級限界**という。また，$a_i - a_{i-1}$ を**級間**

† 実際に度数分布表を作成する際は必ず上記の手順に従う必要はなく，度数分布表がわかりやすくなるように，級間隔や級限界を四捨五入するなどして見やすい数値に調整するとよい（例えば表 1.1 では，$m = 145$，$M = 172$ なので，$a_0 = 145$，$a_k = 172$，$k = 6$ と決めると，級間隔は $(172 - 145)/6 = 4.5$ となるが，四捨五入して級間隔を 5 と調整している。また，それに応じて $a_k = 175$ と調整した）。

隔という†。データがどの階級に入るかを明確にするため，表 1.2 のように級限界に「以上」,「未満」をつけたり，例題 1.1 の表 1.5 のようにデータの値より 1 桁落とした級限界を用いることもある。これを**表 1.3** のように表す。

(4)　階級値を求める。

i 番目の階級 $a_{i-1} \sim a_i$ の階級値 x_i^* は

$$x_i^* = \frac{a_{i-1} + a_i}{2} \quad (i = 1, 2, \cdots, k)$$

で与えられる。

(5)　度数を求める。

各階級に属するデータの個数 f_i $(i = 1, 2, \cdots, k)$ を数え上げる。

表 1.3　度数分布表

階級	階級値	度数
$a_0 \sim a_1$	x_1^*	f_1
$a_1 \sim a_2$	x_2^*	f_2
⋮	⋮	⋮
$a_{i-1} \sim a_i$	x_i^*	f_i
⋮	⋮	⋮
$a_{k-1} \sim a_k$	x_k^*	f_k
合　計	–	n

例題 1.1　**表 1.4** はあるクラスの学生 20 人について，1 ケ月の読書時間〔時間〕を調べた結果である。これをもとに，度数分布表を作成せよ。

表 1.4　学生 20 人の 1 ケ月の読書時間のデータ

20	13	7	2	5	19	1	23	8	17
15	10	18	14	15	9	6	10	9	11

【解答】　**表 1.5** のように，0.5 時間から 25.5 時間の範囲で，階級の数を 5，級間隔を 5 とする。

表 1.5　学生 20 人の 1 ケ月の読書時間の度数分布表

階級	階級値	度数
0.5～5.5	3	3
5.5～10.5	8	7
10.5～15.5	13	5
15.5～20.5	18	4
20.5～25.5	23	1
合　計	–	20

◇

† 級間隔は原則として均一とするが，分布が偏って集中しているときは，均一にしない方が分布の傾向を読み取りやすくなる。例えば，ある市の就業者の年収の分布を調べる場合，1000 万円以下での級間隔は 100 万とし，1000 万円以上の級間隔は 200 万または 300 万などとすることで，より詳しい状況がわかる。

データを視覚的に捉えるためにさまざまな図表が用いられる。横軸に階級をとり，縦軸に各階級の度数を表した柱状のグラフを**ヒストグラム**という。また，i 番目の階級の階級値を x_i^* $(i = 1, 2, \cdots, k)$，度数を f_i とするとき，ヒストグラム上の点 (x_i^*, f_i) を x_i^* の小さい方から順に線分で結んで得られるグラフを**度数折れ線**という。例えば，例題 1.1 で得られた度数分布表からヒストグラム，度数折れ線を作成すると**図 1.1** のようになる。

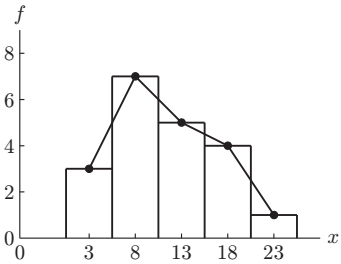

図 1.1　ヒストグラムと度数折れ線

例題 1.2　表 1.2 をもとに，ヒストグラムおよび度数折れ線を作成せよ。

【解答】　表 1.2 から作成したヒストグラム，度数折れ線は**図 1.2** のようになる。

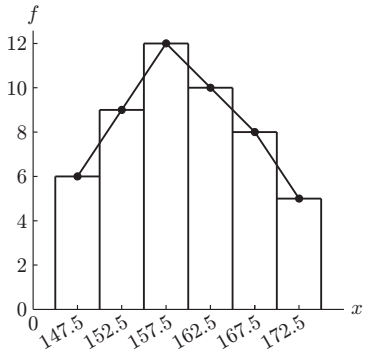

図 1.2　あるクラスの身長のヒストグラムと度数折れ線　◇

各階級の度数を度数の合計で割った値を**相対度数**という。相対度数はその階級の全体に対する割合を表す値である。また，小さい階級値を持つ階級から順に度数を足し合わせたものを**累積度数**，相対度数を足し合わせたものを**累積相対度数**という。これらの情報を度数分布表に加えたものを，それぞれ相対度数分布表，累積度数分布表，累積相対度数分布表などという。相対度数分布表は，データの個数が異なるグループ間の比較をするときなどに用いられる。また，

索　　引

【い】

1 元配置法 *156*
一様分布 *64, 82*
一致推定量 *108*
一般平均 *159, 168*
因　子 *155*

【う】

上側信頼限界 *109*
上側 100α%点 *75*
ウェルチの検定 *141*

【え】

F 分布 *101*

【か】

回帰係数 *23*
回帰直線 *23*
階　級 *2*
階級値 *2*
χ^2 分布 *94*
階　乗 *29*
確　率 *34*
　——の公理 *35*
確率関数 *51*
確率分布 *50*
確率変数 *50*
確率密度関数 *68*
加重平均 *9*
加法定理 *37*
観測度数 *147*

【き】

棄　却 *126*
棄却域 *125*
危険率 *109, 124*
期待値 *53, 69*
期待度数 *147*
帰無仮説 *123*
級間隔 *3*
級間変動 *160*
級限界 *3*
級内変動 *160*
共通部分 *27*
共分散 *19*

【く】

空事象 *32*
空集合 *26*
偶然誤差 *156*
区間推定 *109*
組合せ *29*

【け】

系統誤差 *156*
結果変数 *22*
元 *26*
原因変数 *22*
検出力 *127*
検定統計量 *125*

【こ】

交互作用 *165, 169*
　——による変動 *169*
誤差の許容限度 *112*
誤差変動 *159, 169*
個　体 *1, 84*
根元事象 *32*

【さ】

最小二乗法 *23*
最頻値 *10*
残　差 *23*
散布図 *18*
散布度 *12*

【し】

試　行 *32*
事　象 *32*
指数分布 *83*
下側信頼限界 *109*
実現値 *50*
四分位範囲 *16*
集　合 *26*
修正項 *160, 170*
集　団 *1*
主効果 *159, 168*
順　列 *28*
条件付き確率 *40*
小標本 *114*
乗法定理 *41*

信頼区間 109
信頼係数 109

【す】

水　準 156
推定値 104
推定量 104
数学的確率 35

【せ】

正
　——の完全相関 19
　——の相関 19
正規分布 72
正規母集団 85
制御因子 156
積事象 32
説明変数 22
セル間変動 170
全事象 32
全体集合 27
全変動 159, 169

【そ】

相　関 18
相関係数 21
相関図 18
相対度数 5
相対頻度 36

【た】

第 1 四分位数 16
第 1 種の過誤 127
第 3 四分位数 16
大数の法則 90
第 2 四分位数 16
第 2 種の過誤 127
代表値 8
大標本 113
大標本法 114
対立仮説 124
多元配置法 156

【ち】

中央値 10
抽　出 84
中心極限定理 90
超幾何分布 65

【て】

定性的変数 1
t 分布 97
定量的変数 1
データ 1
適合度の検定 147
点推定 104

【と】

統計的確率 36
統計量 87
特　性 1, 84
独　立 42, 56, 87
独立性の検定 151
度　数 2
度数折れ線 5
度数分布表 2

【に】

二項分布 57
二項母集団 92
2 次元データ 18

【は】

排　反 33
はずれ値 11
範　囲 15
半整数補正 80
反復試行 47
　——の確率 48

【ひ】

ヒストグラム 5
左側検定 124
非復元抽出 85
標準化 77
標準正規分布 73
標準偏差 13, 53, 70
標　本 84
　——の大きさ 84
標本空間 32
標本調査 84
標本標準偏差 87
標本比率 92
標本分散 87
標本分布 88
標本平均 87

【ふ】

負
　——の完全相関 19
　——の相関 19
復元抽出 85
部分集合 27
不偏推定量 105
不偏性 105
不偏分散 87
分割表 150
分　散 13, 53, 70
分散比 162, 172
分散分析法 155

【へ】

平　均 53, 69
平均値 9
平均平方 162, 172
ベイズの定理 45
ベルヌーイ分布 57
偏　差 13
ベン図 27
変　数 1
変　動 159, 169

【ほ】

ポアソン分布 61
補集合 27
母集団 84
母集団分布 85
母　数 85

母標準偏差 85
母比率 91
母分散 85
母平均 85

【み】

右側検定 124

【む】

無限集合 26
無限母集団 84
無作為 35
無作為抽出 84
無作為標本 85
無相関 19

【め】

メディアン 10

【も】

モード 10
目的変数 22

【ゆ】

有意水準 124
有限集合 26
有限母集団 84
有　効 107
有効推定量 107

【よ】

要　素 26
余事象 32

【り】

離散型一様分布 64
離散型確率分布 50
離散型確率変数 50
離散型変数 2
両側検定 124
理論度数 147

【る】

累積相対度数 5
累積度数 5

【れ】

レンジ 15
連続型一様分布 82
連続型確率分布 50
連続型確率変数 50
連続型変数 2
連続補正 80

【わ】

和事象 32
和集合 27

—— 著 者 略 歴 ——

道家　暎幸（どうけ　ひでゆき）
1973年　日本大学大学院理工学研究科修了
　現在　東海大学名誉教授
　　理学博士

伊藤　真吾（いとう　しんご）
2009年　東京理科大学大学院理学研究科修了
　現在　北里大学教授
　　博士（理学）

宮﨑　直（みやざき　ただし）
2010年　東京大学大学院数理科学研究科修了
　現在　北里大学准教授
　　博士（数理科学）

酒井　祐貴子（さかい　ゆきこ）
2007年　東北大学大学院理学研究科修了
2010年　早稲田大学大学院基幹理工学研究科修了
　現在　北里大学講師
　　博士（理学）

はじめての統計学

Introduction to Statistics

2017年 2 月 28 日　初版第 1 刷発行　★

検印省略

著　者　道家暎幸／伊藤真吾／宮﨑直／酒井祐貴子
発行者　株式会社　コロナ社
　　代表者　牛来真也
印刷所　三美印刷株式会社

112–0011　東京都文京区千石 4–46–10
発行所　株式会社　**コロナ社**
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話（03）3941–3131（代）
ホームページ http://www.coronasha.co.jp

ISBN 978–4–339–06113–0　（横尾）　（製本：愛千製本所）
Printed in Japan