

ま え が き

現在は、パソコンの普及により良質な統計処理ソフトが簡単に手に入り、誰もがこれを使うことができます。データを入力するだけでコンピュータは瞬時にいろいろな数値・結果を示してくれます。しかしながら、基本となる統計的考え方や統計的方法の意味がわからないと、結果として表示された数値・結論などを正しく読み取ることはできません。誤った解釈は混乱と間違いを引き起こす因になります。このような観点から、初めて統計学を学ぶ学生諸君には統計的なものの考え方・統計的手法の意味をしっかりと身に付けて欲しいと思います。多少不便はあっても電卓でこつこつ計算を行い、数量からの感触を得ながら統計的な考え方や手法を学ぶことが重要です。

本書は、初めて統計学を学ぶ新入生のための教科書として書かれています。当大学の1年生に対する長年の教育経験から、1年間でこなせると思う内容を厳選しました（しかし、統計学全般から見るとほんの一部です）。大学での1年間の授業は28回から30回くらいですが、3.4節や6.5節などは省略してもよいので、この回数で7章の適合度の検定までは学んで欲しいと思います。8章は医療現場などでよく使われる統計的手法なので、学生諸君は必要に応じて独学して下さい。また、実際に世の中（社会）で使われる統計的方法はその分野ごとに非常に沢山ありますが、本書で学ぶような統計の基礎ができていれば、それらを理解し応用することは難しいことはありません。このような意味で、本書は“統計的考え方の基礎”を身に付けるための教科書であると考えて下さい。

上級年次または社会人になってからさらに統計学を勉強する諸君も出てくることと思いますが、他の本や統計資料・研究論文などが無理なく読めるように、本書で統計学の基礎を固めてくれることを願っています。

2010年2月

大橋 常道

目 次

1. データの整理

1.1 統計的方法とはなにか	1
1.2 度数分布表・ヒストグラム	6
1.3 平均と標準偏差	16
1.4 相関と回帰	24

2. 確率と確率分布

2.1 事象と確率	37
2.2 確率変数と確率分布	53

3. 二項分布と正規分布

3.1 二項分布	66
3.2 正規分布	70
3.3 二項分布の正規近似	75
3.4 ポアソン分布	80

4. 標本分布

4.1 不偏推定量	86
4.2 標本平均 \bar{X} の分布	93

5. 推 定

5.1 母平均 μ の推定	99
5.2 二項母集団の割合 p の推定	107

6. 検 定

6.1 仮説検定とは	110
6.2 母平均 μ の検定	114
6.3 二項母集団の割合 p の検定	118
6.4 2つの母平均の差の検定	121
6.5 2つの割合の差の検定	128

7. カイ2乗検定

7.1 適合度の検定	132
7.2 分割表による独立性の検定	138
7.3 等分散の検定	143

8. 分布型によらない検定

8.1 中央値の検定	149
8.2 ウィルコクソンの順位和検定	151
8.3 ウィルコクソンの符号つき順位和検定	155

付 録	159
引用・参考文献	172
問 の 答	173
問 題 の 答	176
索 引	184

1

データの整理

1.1 統計的方法とはなにか

いまここに、1つの大きな集団（人の集団でも物の集団でもよい）があったとする。集団をつくる個体数は多いと考えてよいので、各個体のある特性を調べるにしても全部の個体を調べることは困難である。このようなとき、この集団からいくつかのデータ（個体を特徴づける数値や性質など。観測値ともいう）をとり、それらを分析・考察することにより、集団全体の特徴や規則性について何らかの結論を出すことは可能だろうか？

人間は長い歴史の中で、このような社会現象の中の集団を扱う方法を確立してきたので当然答を出すことは可能である。ここで、近代統計学の基礎を築いたとされている3人の統計学者の成し遂げた仕事などを以下に記す。

ケトレー（Lambert-A Quetelet, 1796～1874, ベルギー）：近代統計学の祖。社会現象に自然科学の計量的方法および確率論を適用し、統計の基礎を社会現象の合法則性に求めた。統計学を学問として誕生させ、現代統計学への重要な第一歩を踏み出した。

ピアソン（K.Pearson, 1857～1936, イギリス）：大量データをまとめる統計的記述の面を精密化し、統計学を数理的傾向の強いものにした。ピアソンの統計学はいまでは“記述統計学”と呼ばれていて、現在用いられている種々の統計的概念を確立した。

フィッシャー（Ronald.A.Fisher, 1890～1962, イギリス）：最尤推定量の

概念を導入した。実験計画法および分散分析法を開発し、統計学の適用範囲を実験科学の分野にも広げた。母集団と標本の区別を明確にし、現在の推測統計学の方法を確立した。

統計学とは、集団の現象を数量的に観察することにより、その集団がもつ性質や規則性を見つけ出すための方法論を研究する学問であるといえる。

データとは、人間の集団ならば、身長、体重、血圧などを表す数値データおよび血液型 (O, A, B, AB) や個人の性格・特徴 (温厚, 我慢強い, 怒りっぽい など) などを表す質的データの両方を含む。身長、体重、血圧などは一般に**変数** (variable) と呼ばれ、 X, Y, Z など大文字で表す。

現代の複雑で変化の激しい社会活動の中で見られる一見不規則で予測できないような現象でも、いくつかの変数の集団的データおよび定期的な観察を整理・分析することにより、ある規則性や特性が見出されることがある。このことにより、われわれは集団に対してある種の結論を下すことが可能になるのである。

データの源泉としての集団を**母集団** (population) と呼び、母集団を構成する1つ1つを**個体**と呼ぶ。母集団からいくつかの個体を選びデータをとることを**標本抽出**という。得られたデータは1つの変数の実現値と考える。また、データの個数を**標本の大きさ** (size) という。集められたデータは母集団をある程度正確に反映していなければならないので、標本抽出は偏りのない方法、すなわち**無作為** (ランダム) でなければならない。

定義 1.1 (無作為抽出 (random sampling)) 母集団から n 個の個体を抽出するとき、どんな個体も選ばれる機会 (確率) が均等であるような選び方を無作為抽出という。

〔反例〕

- 市長選挙の予測のため、ある政治集会に出かけ行き当たりばったりで 1000 人を選びアンケートをとった際、約 90% の者が特定の政党支持者であった。

これは無作為抽出とはいえない。なぜならば、政治集会なので特定の政

党を支持する者の割合が多くなっている可能性は最初から予想されるからである。無作為抽出のためには、政治集会でない集会、駅頭、市役所や病院などに出かけ、年齢層や男女比にも偏りがないように注意して選ばなければならない。

- 日本人の二十歳の男性の平均身長を予測するために5万人のデータが必要になり、大都市の駅頭でデータを収集した。

これも無作為抽出とはいえない。大都市周辺の男性と田舎の男性とでは環境が異なり、身長に差が生じているかもしれない。また、駅頭でデータをとるというのでは、外出しない人や病気の人を選ぶことはできない。よりよい方法は、全国を二十歳の人口に応じて100等分し、1つの地域から500人を無作為抽出することである。その際、地方自治体の二十歳の男性の名簿を利用することも考えられる。

♡

上の反例で見ると、無作為抽出は一般に非常に難しい作業であるので、実際にデータをとる際は注意を要する。もし、集団が番号づけされていれば、コンピュータで発生させた乱数を利用してサンプリングすることができる。テレビ会社などでは意見を聞くための番号づけされた視聴者の集団をもっているので、視聴率調査などのための無作為抽出は簡単にできるようである。

例 1.1 (乱数表を利用した無作為抽出) 学生数が1000人のある大学の学生自治会は、大学祭の企画について30名の学生から意見を聞きたいと考えている。学生は0番から999番まで番号づけされているので、乱数表を利用して30人を選びたい。ここでは巻末の乱数表を用いて30個の数字を選ぶ。選ぶ際は表のどこから始めてもかまわないのであるが、いま1行目の最初の数から右へ3桁ずつ数字をとっていくと(すでに選んだ数と同じものが出た場合はそれを除いてつぎの数を選ぶ)、

318 76 884 667 284 963 870 214 927 6
872 550 789 887 364 835 957 359 999 704

127 886 420 325 807 132 626 881 315 670

となる。すなわち、これらの番号の学生を選び意見を聞けばよい。♡

さて、母集団から無作為抽出されたデータから、平均や標準偏差を求めたり、度数分布表やグラフを作ったりして、データ全体の特徴が見てすぐわかるようにまとめること（記述統計学）が最初の仕事である。つぎに、問題にしている変数の分布やわかっている性質（理論）などとすでに計算された統計量を利用して、最終的には母集団に対してなんらかの結論を出す（推測統計学）ことである。すなわち、統計学では得られたデータから確率分布や確率を用いて、最大限期待されると考えられる答えを出すのである。このような統計的方法(図 1.1)は、数学や物理でやるような“真実を求める”という決定論的方法とは異なる。推測統計学においては、データの源泉としての変数は一般に確率変数とみなす。

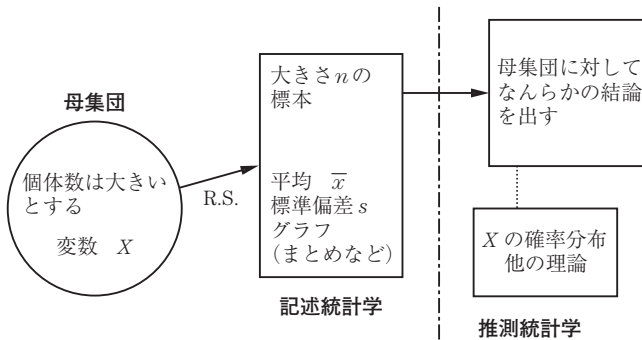


図 1.1 統計的方法

定義 1.2 (確率変数 (random variable)) X が確率変数とは、そのとる値が偶然性に依存していると同時に、それらの値をとる確率が定められていることである。

注意：確率変数の一般的な定義は 2 章で再び与えるが、ここではサンプリングあるいは 1 つの実験のたびごとに異なる値を取り得る変数と理解してかまわない。

例 1.2

- (1) サイコロを 1 回投げる実験で出る目の数を X としたとき、 X は確率変数であり、取り得る値は $1, 2, \dots, 6$ である。また周知のごとく、 X がこれらの 1 つの値をとる確率は $\frac{1}{6}$ なので、つぎのように表現する：

$$P(X = i) = \frac{1}{6} \quad (i = 1, 2, \dots, 6).$$

- (2) ある大学の男子学生の集団で、 Y を体重としたとき、 Y は確率変数である。なぜならば、データを選ぶたびに値は異なり、予測できるものではないからである。取り得る値は、ある区間の任意の値である。変数 Y の確率分布がどのようになるかをここで説明することは難しい問題であるが、2 章で言及する。 ♡

確率変数は図 1.2 の樹形図のように分類される。

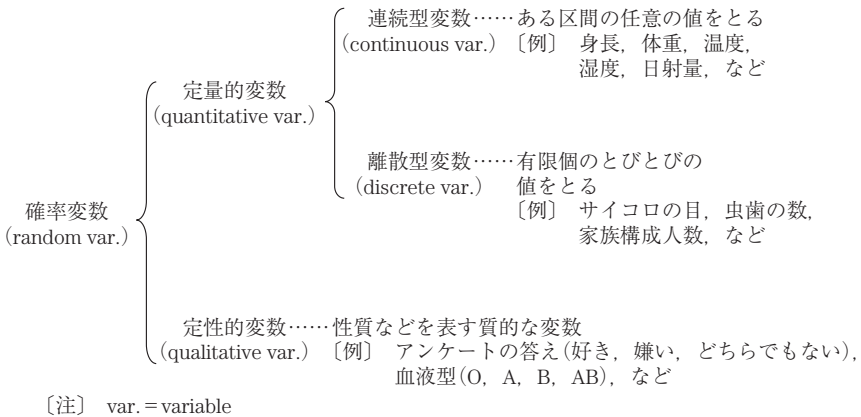


図 1.2 確率変数の分類

問題 1.1

問 1. つぎのデータの抽出法は無作為抽出といえるか。

- (1) A 大学では、学生のアレルギー体質をもつ者の割合を調べるため、登校途中の学生 100 人を選びデータをとった。

6 1. データの整理

- (2) ある都市の勤労者の収入分布を調査するため、電話帳で 10 番目ごとに名前を抽出するという方法で 500 人を選んだ。
- (3) ビタミン C 錠剤を作っている B 製薬会社は、製品の抜き取り検査をするため、1 年を四半期に分けてそれぞれの期間に作られた製品の中から 100 個をランダムに選び、合計 400 個の錠剤を検査した。
- (4) 市長選挙の結果を予測するため、5 つの大きなデパートを選び、その入口でランダムにお客から意見を聞いた。
- 問 2. ある町の年金受給者 600 人には、1 番から 600 番の番号が付けられている。町政についての意見を聞くため、20 人を無作為抽出したい。巻末の乱数表を用いて 20 人を選べ。
- 問 3. つぎの変数 X, Y, Z は確率変数かどうか答えよ。
- (1) サイコロを 3 回投げる実験において、
 X : 1 の目が出る回数, Y : 出た目の和。
- (2) A 大学の平成 20 年 1 月 * 日の選抜入試の受験者の集団に対して、
 X : 英語の得点, Y : 化学の平均点, Z : 欠席者の人数。

1.2 度数分布表・ヒストグラム

どんな目的にしろ、集められたデータは、その数値または記号を眺めているだけでは全体を把握することはできない。この節では、データのまとめ方とグラフ表示について述べる。以下 2 ページにわたるデータ (表 1.1) は、ある年の K 大学 1 年生男子 55 名のデータである。これを基にして説明する。

表 1.1 のデータでは、変数は 3 種類に分けられる：

- (1) 連続型変数 身長, 体重, 靴の大きさ, 母の身長, 父の身長
- (2) 離散型変数 数学, 英語, (これらは高校時代のおおよその得点平均), 劇場回数 (1 年間に映画館やコンサートホールなどへ行く回数)
- (3) 定性的変数 血液型, (いま 1 番の) 関心事, アレルギー体質 (種類)
- さて、劇場に行く回数のような離散型データは特に工夫しなくても自然にまとめられる。表 1.2 と図 1.3 の度数分布表とヒストグラム (度数分布を表す柱状グラフ) を参照せよ。度数分布表の階級 (class) の大きさは、0 から 6 以

表 1.1 200X 年, 男子学生データ

No.	性別	血液型	身長	体重	靴大きさ	母の身長	父の身長	数学	英語	劇場回数	関心事	アレルギー	アレルギー種類
1	M	O	171.2	65.0	26.0	162	180	80	50	5	自分自身の生き方	なし	
2	M	O	170.0	65.0	25.5	160	175	80	70	3	自分自身の生き方	なし	
3	M	A	172.0	61.0	26.5	155	165	50	50	2	スポーツ・娯楽	なし	
4	M	O	168.0	62.0	26.0	151	175	85	90	0	生活費(金銭)	あり	花粉症
5	M	B	160.0	55.0	26.0	140	165	80	80	0	スポーツ・娯楽	なし	
6	M	A	174.0	68.0	27.0	158	160	85	50	0	スポーツ・娯楽	なし	
7	M	O	171.0	53.0	26.5	150	170	30	70	0	スポーツ・娯楽	あり	花粉症
8	M	AB	172.0	63.0	26.5	160	165	50	50	4	スポーツ・娯楽	なし	
9	M	B	175.0	68.0	27.0	165	170	60	80	1	生活費(金銭)	なし	
10	M	B	162.0	54.0	26.5	151	170	60	70	2	スポーツ・娯楽	なし	
11	M	O	165.5	49.5	26.0	148	164	65	80	4	スポーツ・娯楽	あり	花粉症
12	M	A	180.0	70.0	26.0	155	180	80	60	12	スポーツ・娯楽	なし	
13	M	B	175.9	77.5	28.0	158	175	40	40	0	自分自身の生き方	なし	
14	M	A	169.0	85.0	26.5	151	165	80	70	0	自分自身の生き方	あり	花粉症
15	M	AB	180.0	58.0	27.5	165	170	76	28	1	スポーツ・娯楽	あり	アトピー・性皮膚炎
16	M	O	163.0	74.0	26.0	155	172	80	75	2	生活費(金銭)	なし	
17	M	A	176.0	65.0	26.5	152	175	80	80	0	スポーツ・娯楽	なし	
18	M	O	163.0	51.0	25.0	151	167	70	60	2	自分自身の生き方	なし	
19	M	O	180.0	58.0	27.0	156	170	85	40	2	スポーツ・娯楽	なし	
20	M	O	177.0	66.0	27.0	164	179	40	60	1	友人関係	なし	
21	M	AB	181.0	60.0	28.0	164	170	30	55	2	生活費(金銭)	なし	
22	M	O	177.3	65.0	27.0	152	174	30	50	3	社会情勢(政治・経済など)	あり	花粉症
23	M	A	186.0	70.0	29.0	160	170	45	25	1	自分自身の生き方	あり	花粉症
24	M	B	178.0	65.0	27.5	155	170	60	30	2	自分自身の生き方	あり	アトピー・性皮膚炎, 喘息
25	M	AB	157.0	60.0	25.5	150	150	75	85	0	社会情勢(政治・経済など)	なし	
26	M	O	165.0	50.0	26.5	150	160	50	60	2	友人関係	あり	花粉症, 鼻炎
27	M	B	167.0	70.0	26.5	160	171	62	47	1	スポーツ・娯楽	なし	
28	M	AB	175.0	60.0	27.0	155	160	50	40	2	自分自身の生き方	なし	

索引

【あ行】	【く】	【す】
一様分布 64	空事象 38	推測統計学 4
ウィルコクソンの順位和	区間推定 99	推定値の誤差 101
検定 152	組合せ 47	スチューデント
ウェルチの t 検定 147		——の t 変数 96
重みつき平均 21	【け】	【せ】
【か】	経験的確率 42	正規確率紙 15
回帰 24	決定係数 36	正規近似 94, 108
回帰直線 32	ケトラー 1	正規分布 70
階級 6	検定 110	——の近似 79
階級値 10	【こ】	正規母集団 93
カイ 2 乗分布 132	ゴセット 81	正の相関 28
確率 39	個体 2	積事象 38
——の木 50	根元事象 37	線形補間 74
確率分布 54	【さ】	全事象 37
確率変数 4	最小 2 乗直線 32	【そ】
確率密度関数 59	最小 2 乗法 31	相関 24
仮説 110	再生性の定理 90	相関係数 27
仮説検定 111	採択 111	相関関 25
片側検定 116	算術的確率 41	相対度数 9
加法定理 43	散布図 25	【た】
完全相関 28	【し】	対応のある場合の検定 127
ガンマ関数 97	試行 37	対応のない場合の検定 127
【き】	事象 37	対数正規確率紙 15
幾何学的確率 42	四分位範囲 22	大標本法 103, 108, 129
棄却 111	自由度 96	対立仮説 110
棄却域 112	順位和 153	第 1 種の過誤 112
記述統計学 4	順列 47	第 2 種の過誤 112
期待値 54, 61	条件つき確率 44	単一事象 37
期待度数 134, 139	小標本 103, 149	
帰無仮説 115	乗法定理 45	
共分散 26		

【ち～て】

中央値 22, 149
 —の検定 149
 中心極限定理 90
 対標本モデル 155
 定性的変数 6
 適合度の検定 134
 データ 1
 点推定 99

【と】

統計学 2
 統計的方法 4
 同時確率密度関数 62
 等分散の検定 143
 独立 45, 56
 独立試行 65
 独立事象の乗法定理 45
 独立性の検定 140
 度数分布表 6

【に、の】

二項係数 48
 二項分布 67
 二項変数 66
 二項母集団 66
 ノンパラメトリックな方法 149

【は】

排反 39
 範囲 17

【ひ】

ピアソン 1

ヒストグラム 6
 標準化の公式 72
 標準正規分布 71
 標準偏差 18, 54, 61
 標本 87
 —の大きさ 2
 標本空間 37
 標本点 37
 標本標準偏差 96
 標本分散 87
 標本平均 87

【ふ】

フィッシャー 1
 —の直接計算法 142
 複合事象 37
 符号検定 151
 符号つき順位和 156
 負の相関 28
 不偏推定量 87
 分割表 138
 分散 18, 54, 61
 分布関数 59

【へ】

平均 16, 54
 ベイズの公式 49
 ベータ関数 144
 ベルヌーイ試行 65
 偏差値 63
 ベン図 38
 変数 2

【ほ】

ポアソン分布 80
 母集団 2

母数 86
 母分散 86
 母平均 86
 —の差の検定 123

【ま行】

マン・ホイットニー検定 152
 無作為抽出 2
 無作為標本 87
 無相関 28
 モード 21

【ゆ、よ】

有意水準 114
 有意である 115
 有意でない 115
 有効推定量 92
 有効数字 30
 余事象 38

【ら行】

乱数表 3
 離散型確率変数 53
 離散型変数 6
 両側検定 116
 累積度数折れ線グラフ 10
 連続型確率変数 59
 連続型変数 6
 連続補正 77

【わ】

和事象 38
 割合の差の検定 129

【F, H】

F 分布 143
 H_0 曲線 112

H_1 曲線 112

【数字】

2 変数データ 24

3 つの事象が独立 51
 95 %信頼区間 101

— 著者略歴 —

大橋 常道 (おおはし つねみち)
1969年 東京理科大学理学部応用数学科卒業
1972年 東京理科大学大学院修士課程修了
(数学専攻)
1976年 青山学院大学情報科学研究所助手
1980年 北里大学講師 (教養部)
2004年 北里大学教授 (一般教育部)
2012年 北里大学退職

山下 登茂紀 (やました ともき)
1998年 神戸大学理学部数学科卒業
2001年 神戸大学大学院博士前期課程修了
(数学専攻)
2004年 神戸大学大学院博士後期課程修了
(構造科学専攻)
博士 (理学)
2004年 慶應義塾大学理工学部 COE 博士研究員
2005年 朝日大学講師 (歯学部)
2008年 北里大学講師 (一般教育部)
2011年 近畿大学講師 (理工学部)
現在に至る

谷口 哲也 (たにくち てつや)
1992年 東京理科大学理学部第一部物理学科
卒業
1994年 東北大学大学院博士前期課程修了
(数学専攻)
1999年 東北大学大学院博士後期課程修了
(数学専攻)
博士 (理学)
2003年 東北大学大学院理学研究科数学専攻
COE フェロワー
2004年 北里大学講師 (一般教育部)
2010年 北里大学准教授 (一般教育部)
現在に至る

初学者にやさしい統計学

Statistics — A Kind Introduction for Beginners —

© Ohashi, Taniguchi, Yamashita 2010

2010年4月16日 初版第1刷発行

2015年3月20日 初版第6刷発行

検印省略

著者 大橋 常道
谷口 哲也
山下 登茂紀

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 978-4-339-06090-4 (金) (製本：愛千製本所)

Printed in Japan



本書のコピー、スキャン、デジタル化等の無断複製・転載は著作権法上での例外を除き禁じられております。購入者以外の第三者による本書の電子データ化及び電子書籍化は、いかなる場合も認めておりません。

落丁・乱丁本はお取替えいたします