基礎統計学

基礎統計学

新井宏尚著

博文社

基　礎　統　計　学

昭和55年7月20日　発　行

検印
省略

著　者　新井宏尚

発行者　鈴木貞義

発行所　株式会社　文化書房博文社

112　東京都文京区目白台1～9～9
電話　東京　(947)2034(代)
振替　東京　8-8 6 9 5 5番

印　刷　　多田印刷株式会社

製　本　　風林社塚越製本

落丁本・乱丁本はお取替えいたします。

0030—40950—7361

はしがき

はじめて統計学を学習しようとする方々に対してその入門書を執筆することは決して容易でない。その主な原因は統計学特有の諸概念（確率変数，尤度，統計量など）が初学者にとって中々理解されにくいからであろう。

本書の執筆者，新井宏尚氏は早稲田大学大学院修士課程終了後，青山学院大学大学院博士課程に入り，ここで，小生の指導をうける形で，数理統計学を研究された若い新進の研究者であり，一方城西大学，中央大学，青山学院大学などで，統計学の講義を通じて学生を指導されてきた。

その豊かな経験をもとに，初学者向けのテキストとして書かれてあり，著者自身の苦心されたところ，及び小生が学生指導にあたってとくに注意すべき事項としてあげたところが，よくかかれている。

巷間，統計学の入門書として書かれたものは大変多い。中には明らかに誤りの散見される入門書も見受けられるようである。

ところで，小生は日頃，次のように考えている。「学術論文では少しの誤りや論旨の不明確さも許されない。しかし教科書では，少しくらい誤りがあってもよい。何故ならこれをもとに学ぶ人，講義担当者が，誤りに気付いて，さらに正しい理解を深めることが必要であり，全く寸分のスキもない平板なテキストでは，そのまま鵜呑みにされて，重要な点に対して必ずしも正しい理解ができず，いわゆる学問が身につかない」

小生がはじめて統計学の講義に接したのは昭和18年頃である。その後も何人かの先生の講義をうけたり，同学の友人と研究会をもったりして，統計学に接してきた。そのたびに，

(1) 確率変数と数学の変数との本当のちがいは何であるか？

(2) 母集団とその一部である標本という対決すべきものもないのに，何故標

本空間というものが確率論の出発点にかかれるのか？

(3) 仮説検定とは仮説の統計的検定か，統計的仮説の検定か？　もし後者なら，統計的仮説と非統計的仮説とはどこがどうちがうのか？

(4) 期待値と平均値とは同じか，ちがうかちがうなら，どこがどうちがうか？

(5) 直線や曲線（指数曲線や対数曲線など）をあてはめるのに何故1階微分して0とおくだけでよいのか。この必要条件以外に十分条件を何故検討しないのか？

といった素朴な，かつ重要な問題が提起された。

本書でもこうした諸問題のすべてに解答を与えている訳ではないがそのヒントが述べられているようである。

著者が具体例を中心に「統計学の考え方がどうすすめられるのか」を苦心してまとめられたこの労作を一通り拝見したが，これを初めて統計学を学ぶ方々にとって十分役に立つものであることを確信したので，ここに小生流の推薦の辞をかくこととした。

内容についての御意見や御批判は新井宏尚氏または小生あて頂ければ幸甚である。それによって今後さらに加筆し補足することを考慮している。

昭和55年6月　　鈴　木　栄　一

目　次

第1章　資料*1の整理(I)

1-1　はじめに

我々は日常生活では種々の情報に囲まれていると良く言われる。例えば，新聞・雑誌・テレビなどにより，日々の気温，交通事故死亡者数，前年度比何%の物価上昇率とか様々の数量的資料に接する機会も多く，その処理の必要性も生じて来た。一例としてそのような資料をあげる。表1-1，図1-1は朝日新聞に載っていた(1979.12月)表とグラフであり，表は1979年度の各都市の月平均気温の資料(データ，標本値)である。図1-1は1976年から1979年にかけての日本上空5500m付近の気圧の変化がどのような状態にあるかをわづか39個

表 1-1

		1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	平年気温
平均気温(上段)と平年差	札幌	−5.3	−2.6	−0.6	4.6	10.9	17.5	19.4	22.1	16.9	12.5	4.7	7.8
		−0.2	1.8	0.0	−1.5	−0.9	1.8	−0.8	0.4	0.0	2.1	1.0	
	仙台	2.7	4.1	4.9	9.3	14.9	21.4	22.0	24.9	20.6	16.5	10.5	11.6
		2.1	3.2	1.1	−0.3	0.4	3.2	0.3	0.9	0.6	2.5	2.1	
	東京	6.6	8.4	9.9	13.9	18.7	24.4	25.2	27.4	24.1	19.6	14.3	15.0
		2.5	3.6	2.0	0.4	0.7	3.1	−0.4	0.7	1.1	2.7	2.6	
	名古屋	5.6	7.2	8.2	13.0	18.1	23.6	25.2	27.4	23.8	18.5	12.6	14.7
		2.4	3.4	1.2	−0.1	0.3	2.0	0.0	0.5	0.9	1.9	1.8	
	新潟	3.5	4.9	5.8	10.1	15.3	22.4	23.8	25.7	21.7	17.3	11.7	13.0
		1.7	3.0	1.1	−0.3	−0.4	2.5	−0.5	−0.2	0.3	1.9	1.9	
	大阪	7.2	8.5	9.2	13.9	19.0	24.5	26.5	29.1	24.7	19.7	13.7	15.6
		2.7	3.6	1.2	0.0	0.4	2.0	−0.5	1.1	0.8	2.1	0.5	
	福岡	7.7	8.0	9.9	14.0	18.1	23.6	26.3	27.9	24.6	19.0	12.5	15.7
		2.4	2.0	0.9	0.1	0.0	1.9	−0.8	0.7	1.3	1.7	0.0	

(朝日新聞より)

*1　資料とは確率変数の実現値であり，計量値(重さ，長さなど)計数値(人口，不良品数など)と呼ばれることもある。

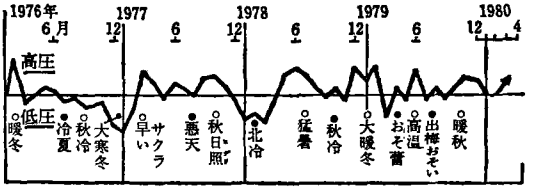


図 1-1　日本上空 5,500 m 付近の気圧の変化に見る気候変動　(朝日新聞刊)

の観測値が表わしている。これら図表はデータがその源泉(母集団*2)について役立つ情報を提供することを教えてくれる。このような資料の集まりを標本と言い，標本からその源泉である母集団に関して有用な情報を引き出すこと，つまり既知の標本から未知の状態にある母集団について何らかの結論を与えること(帰納的役割)が統計的方法の重要な役割であり，この部門を推測統計学と言う。また，データを収集したり，要約したりする役目を果すのが記述統計学である。

1-2　資料の分類とグラフ化

表 1-1 の各都市の月平均気温を示す数値は何度何分と記録されているが，より精度の高い記録計があるならばより詳細な記録を得ることができよう。気温，体重，身長，時間間隔等の測定値，観測値はある区間内の任意の値をとることができる。このような変量を連続型変量または連続型変数と言う。ただし，測定器具の精度にも限度があるのでいくらでも詳細に記録できるわけではないが，連続型変量とみなす。他方，世帯内の子供の数，交通事故数，わが国の総人口，ある地域内の住宅数とかのような変量を離散型変量，変数と言う。この例として，表 1-2，図 1-2 を載せておこう。表 1-2 は犯罪検挙人員のデータである。図 1-2 はこれをグラフ化したもので円グラフと言われる。

表 1-2 で使用されている構成比は総数で項目内の数を割ったものである。図

*2　母集団は様々な現象に対して定義されるが JISZ 8101 品質管理用語では
i)　調査，研究の対象となる特性をもつすべてのものの集団
ii)　試料やデータにより処置をとろうとする集団と定義されている。

表 1-2 主要罪名別刑法犯検挙人員*3

(昭和51年，52年)

罪名	51年		52年		前年差	
	検挙人員	構成比	検挙人員	構成比	実数	増減率
総数	830,679	100.0	822,218	100.0	− 8,461	− 1.0
窃盗	201,932	24.3	207,064	25.2	+ 5,132	+ 2.5
強盗	939	0.1	814	0.1	− 125	−13.3
強盗致死傷，強盗強姦	1,106	0.1	1,012	0.1	− 94	− 8.5
詐欺	15,918	1.9	15,665	1.9	− 253	− 1.6
恐喝	10,686	1.3	9,660	1.2	− 1,026	− 9.6
横領	9,904	1.2	12,375	1.5	+ 2,471	+24.9
殺人	2,113	0.3	1,988	0.2	− 125	− 5.9
傷害・同致死	40,590	4.9	40,730	5.0	+ 140	+ 0.3
暴行	26,368	3.2	25,781	3.1	− 587	− 2.2
強姦・同致死傷	3,394	0.4	3,046	0.4	− 348	−10.3
放火	876	0.1	921	0.1	+ 45	+ 5.1
業務上(重)過失致死傷	473,638	57.0	461,353	56.1	−12,285	− 2.6
その他	43,215	5.2	41,809	5.1	−1,406	− 3.3

注 警察庁の統計による。

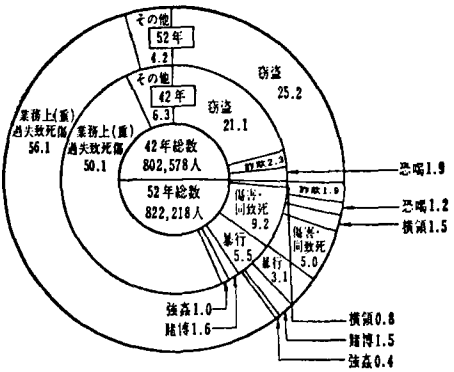
1-2の円グラフは，表の構成がどのようになっているのかを見ればすぐにわかるという利点がある。図1-1のグラフを時系列グラフというが，これは変量が時間の経過に伴ってどのように変化するかが一目で判明するという長所がある。他には絵グラフなどがあるが最も代表的グラフは次に述べる度数グラフである。

いま統計学のテストを72人の学生に対して行ったところ(表1-3)，(表1-4)が得られたとしよう。

表1-3は72個の素点を羅列しており，表1-4は10点間隔で整理したものである。この場合の度数とは各間隔に含まれる人数を示している。度数を示すマークはわかりやすいならばどのようなマークでも差しつかえない。

*3 昭和53年度犯罪白書より引用

① 刑法犯



② 業過を除く刑法犯

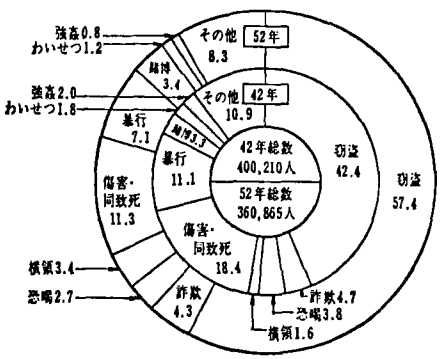


図 1-2　主要罪名別刑法犯検挙人員の構成比

(昭和42年，52年)

表1-4を主としてまづ度数グラフ(ヒストグラム)を作成してみよう。図1-3が度数グラフである。また図1-4が累積度数グラフと累積相対度数グラフである。データのもつ特性によりヒストグラムの形状も様々であるが図1-3のように峰(または山)が1つ(単峰型)の場合もあれば，複数個の峰をもつ度数分布もある。

資料が与えられた場合，度数(分布)グラフの作成手順は通常次のようなステ

表 1-3 72人の学生の得点結果

93,	25,	15,	0,	65,	70,	80,	50
63,	48,	39,	42,	67,	44,	18,	93
84,	59,	58,	62,	74,	68,	54,	11
24,	29,	87,	66,	73,	55,	34,	77
63,	69,	54,	33,	47,	71,	83,	64
59,	27,	73,	41,	58,	55,	39,	41
34,	9,	18,	88,	68,	73,	79,	61
91,	99,	57,	69,	67,	71,	89,	88
67,	44,	81,	78,	74,	51,	68,	74

表 1-4

点　数	度数(人数)	累積度数	相対累積度数
0以上～10未満	2人 丁	2	0.03
10～20	3人 正	5	0.07
20～30	3人 正	8	0.11
30～40	5人 正	13	0.19
40～50	7人 正丁	20	0.29
50～60	11人 正正一	31	0.44
60～70	15人 正正正	46	0.66
70～80	12人 正正丁	58	0.83
80～90	8人 正下	66	0.94
90～100	4人 正	70	1.00

ップである。

ステップ1　最大値と最小値をもとめ，範囲を決める。例では0点と99点であるので範囲は99-0となる。

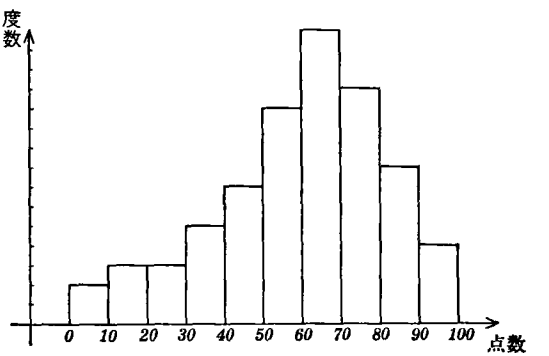


図 1-3

ステップ2　クラスの数を決める。例では10クラスである。通常クラスの数は10～15前後であろう。そして，各クラスに入るデータ数を数え度数を求めればよい。クラスの数はデータの数により左右されるので常に決まったクラス数があるわけでないが，データ数が50～100ならば10前後のクラスで十分であろう。データの数が少ないとグラフの形がわかりにくいのでグラフを作成するときの資料にもよるが通常少なくとも30個，できれば50

個以上が望まれる。

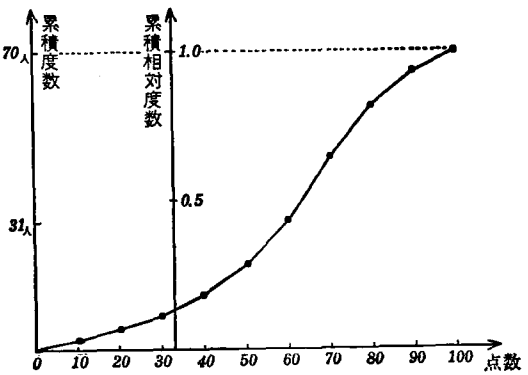


図 1-4

1-3　資料の数量的表示

前節で簡単ではあるが資料のグラフ化について述べた。しかし，視覚によるだけでは資料の持っている情報を十分に利用しているとは言えないので，資料の持っている特性を数量的尺度(測度)で示す必要がある。測度としては中心的傾向を示す測度，散らばり具合を示す測度があるので以下順次それらを見てみよう。

1-3-1　中心的傾向を示す測度

(a)　算術平均

n 個の資料がある場合，それらを $x_1, x_2, \cdots x_n$ で表わすとき，算術平均は次式で与えられる。

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad (1\cdot1)$$

例1　表1-1から札幌，仙台，東京，名古屋，新潟，大阪，福岡の1月の月平均気温をそれぞれ $x_1, x_2, \cdots, x_7$ とすれば

$$\bar{x}=(-5.3+2.7+6.6+5.6+3.5+7.2+7.7)\div 7=4$$

となる。また平均値を求める場合の簡便計算方法として，$y_i=x_i-x_0$ (x_0 とはある一定の値で通常平均値に近い値に見当をつけ，それを x_0 とするので仮平均と言われる.) と置き換えて，$\bar{x}$ を求める。

$$\bar{x}=x_0+\frac{1}{n}\sum_{i=1}^{n} y_i \qquad (1\cdot 2)$$

上の例の場合，$x_0=4$ とおけば y_i の値は

$y_1=-9.3$, $y_2=-1.3$, $y_3=2.6$, $y_4=1.6$, $y_5=-0.5$, $y_6=3.2$, $y_7=3.7$

となる。従い

$\bar{x}=4+(-9.3-1.3+2.6+1.6-0.5+3.2+3.7)/7=4$ を得る。

(b) 算術平均(度数がある場合)

度数がある場合には度数を f_i，階級値を x_i とすれば $\bar{x}$ は

$$\bar{x}=\frac{x_1\cdot f_1+x_2\cdot f_2+\cdots+x_n\cdot f_n}{N}=\frac{1}{N}\sum_{i=1}^{n} x_i f_i \qquad (1\cdot 3)$$

で求められる。ただし，$N=f_1+f_2+\cdots+f_n=\sum_{i=1}^{n} f_i$ とする。階級値 x_i とはクラス幅の代表値のことである。

表 1-5

x_i	f_i	$x_i\cdot f_i$	u_i	$u_i f_i$
5	2	10	−5	−10
15	3	45	−4	−12
25	3	75	−3	−9
35	5	175	−2	−10
45	7	315	−1	−7
55	11	605	0	0
65	15	975	+1	+15
75	12	900	+2	+24
85	8	680	+3	+24
95	4	380	+4	+16
	計 70	計4160		計+31

例 2 表 1-4 から平均を求めてみよう。表 1-4 を整理したものが表 1-5 である。

表 1-5 及び式 (1·3) より $\bar{x}=\sum_{i=1}^{n} x_i f_i/N=59.4$ がえられる。また，クラス分けされた資料の平均を求める場合の簡便法として符号変数 (coding variable) u による計算方法がある。

仮平均値に対応する u_i を $u_i=0$ とおき，階級のクラス幅を c とするとき，

*4 $\bar{x}$ は変量と同一次元で表現されなければいけない。

x_i と u_i と c の関係は次のようになる。

$$x_i = cu_i + x_0, \quad c=10, \quad x_0=55 \tag{1·4}$$

x_0 は $u_i=0$ に対応する値である。したがって平均 x は

$$x = c\bar{u} + x_0 = 10\times 0.44 + 55 = 59.4 \tag{1·5}$$

となる。ただし，$\bar{u}=\dfrac{1}{\sum f_i}\sum u_i f_i$ とする。

(c)　メディアン(中位数，中央値)

資料を大きさの順に並べたとき，ちょうど中央にくる資料があるならば，その資料をメディアンという。資料の数が偶数ならば中央の2つの相加平均がメディアンになる。記号は M_e で記される場合が多い。

例3　50kg, 60kg, 65kg, 63kg, 59kg, 58kg, 57kg, 61kg のメディアンを求めよ。

50, 57, 58, 58, 60, 61, 63, 65 となり偶数個であるから，$(59+60)\div 2=59.5$ となる。

メディアンの長所は最後の階級値が「～以上を全て含む」という形になっている場合，分布の両端で極端な値が出た場合にもそれに影響されない点である。度数分布表から M_e を求める場合 n が偶数奇数を問わず $\dfrac{n}{2}$ となる x_i を M_e と決める場合が多い。累積度数を使用する場合には次のようにすれば良い。いま i 番目の階級の度数を含めると累積度数が $\dfrac{n}{2}$ を越え，含めなければ $\dfrac{n}{2}$ に達しないとする。このとき $(i-1)$ 番目までの累積度数を F_{i-1}，階級 i の度数 f_i，$(i-1)$ 番目のクラス間隔の上端を ll_{i-1} とすると，メディアンは

$$M_e = ll_{i-1} + c\cdot\frac{\dfrac{n}{2}-F_{i-1}}{f_i} \tag{1·6}$$

となる。ただし c はクラス間隔である。

(d)　モード(最頻値)

最大の度数を持つ測定値が1つあるとき，その測定値をモード(最頻値)という，度数分布のようにクラス分けされた資料ではモードは次式で計算される。